Guidelines on the Verification of Hydrological Forecasts

2025 edition

WEATHER CLIMATE WATER

WORLD METEOROLOGICAL ORGANIZATION

WMO-No. 1364

WMO-No. 1364

Chair, Publications Board
World Meteorological Organization (WMO)
7 bis, avenue de la Paix
P.O. Box 2300
CH-1211 Geneva 2, Switzerland

Tel.: +41 (0) 22 730 84 03
Email: publications@wmo.int

ISBN 978-92-63-11364-0

CONTENTS

FOREWORD

In the complex and multifaceted field of hydrology, the verification of streamflow forecasts plays a crucial role in understanding and predicting water movement within river basins.

Flood preparedness and emergency response often prioritize the verification of river stage or flow height forecasts, given their direct impact on critical thresholds and timing of flow peaks. This publication provides comprehensive guidelines focused on the verification of forecasted streamflow, or discharge, at designated locations. It aims to provide detailed methodologies and considerations for streamflow forecast verification.

Verification of forecasts plays a vital role by providing several key benefits: it provides operational insight by highlighting the strengths, weaknesses and uncertainties of the forecasts and their systems; it guides system enhancements; and ultimately it enables more informed and effective decision-making.

Highlighting the uncertainty of forecasts is particularly important: it must be remembered that all forecasts, whether deterministic or probabilistic, contain uncertainty and errors that form an essential part of the forecast information. Verification of forecasting is therefore needed to increase the understanding by forecasters of the quality of these forecasts, and to improve the way the forecasts are communicated to end users.

Through these functions, verification supports the continuous improvement and reliability of forecasting systems.

While the subject is technical, the proposed guidelines are designed to introduce verification concepts to beginners and assist practitioners in evaluating their operational system. They should help the community select appropriate verification metrics and tools to assess the value of their operational forecasts.

Recognizing the diversity of hydrological forecast users, their varying risk perceptions, and the differing predictive skills across hydroclimatological regions, this publication avoids a one-size-fits-all approach. Instead, it provides general guidance, key points, case studies and examples to help users tailor verification tasks to their specific needs.

We hope this publication serves as a valuable resource for professionals dedicated to improving the accuracy and reliability of hydrological forecasts, ultimately contributing to more informed decision-making and enhanced flood preparedness, and supporting the Early Warnings for All (EW4All) initiative.

Prof. Celeste Saulo
Secretary-General

PREFACE

The verification of hydrological forecasts is critical for retrospectively assessing the performance of operational flood forecasting systems. This topic holds significant interest for members of the hydrological forecasting community as it serves as a benchmarking for evaluating the quality of their forecasts. Though specific to the hydrological sector, these guidelines provide an additional element contributing to the Early Warnings for All (EW4All) initiative, allowing Members to advance in their hydrological forecasting capabilities.

This activity was initiated in 2018 under the former Working Group on Hydrological Services (WGHS) of Regional Association II, with S. V. Borsch, A. V. Khristoforov, and E. A. Leonteva, from the Russian Federation, who embraced the initial effort of drafting an overview document. Following the WMO restructuring, the Standing Committee on Hydrological Services (SC-HYD) took over the development of these guidelines.

As President of the Commission for Weather, Climate, Hydrological, Marine and Related Environmental Services and Applications (SERCOM), I extend my deepest gratitude to all contributors involved in developing these guidelines, especially the authors, Julie Demargne (HYDRIS hydrologie), Jan Simon Verkade (Deltares) and Dong-Jun Seo (University of Texas, Arlington). The commitment and leadership of SC-HYD members Reggina Cabrera and Paolo Reggiani in supervising the drafting process have been instrumental during this process.

I would like to also thank James Bennett, Durga Lal Shrestha and David Robertson, who contributed case studies and examples.

The authors would like to thank the United States National Weather Service Middle Atlantic River Forecast Center in State College, Pennsylvania, West Gulf River Forecast Center in Fort Worth, Texas, and Office of Water Prediction in Silver Spring, Maryland, for making available the various datasets used in this report. Some contributions to this publication are based on materials supported in part by the National Oceanic and Atmospheric Administration (NOAA) Climate Program Office under grant NA15OAR4310109 and Joint Technology Transfer Initiative Program under grants NA17OAR4590174, NA17OAR4590184 and NA16OAR4590232, the National Weather Service Cooperative Program for Operational Meteorology, Education, and Training Program Subaward No. SUBAWD000020 and the National Science Foundation under grant CyberSEES-1442735.

These guidelines benefited from technical reviews by the following experts:

- Thomas Pagano (Australian Bureau of Meteorology)
- James Brown (Hydrologic Solutions Limited, United Kingdom of Great Britain and Northern Ireland), for review of the EVS computational example
- Thomas Nipen (Norwegian Meteorological Institute), for help with and review of the verif computational example
- Francesco Laio (Department of Environmental Engineering, Land and Infrastructure Management (DIATI), Polytechnic of Turin, Italy)
- Marc Philippart, Annette Zijderveld (Rijkswaterstaat, Kingdom of the Netherlands) and Maarten Smoorenburg (Deltares, Kingdom of the Netherlands) for reviewing the Rijkswaterstaat operational verification case study
- Justin Robinson and colleagues (Australian Bureau of Meteorology) for reviewing the section about the Bureau's Performance Analysis Tool
- Dominic Roussel and Simon LaChance-Cloutier (Québec Ministry of the Environment, the Fight Against Climate Change, Wildlife and Parks, Canada) for reviewing the section about the Système de Prévision Hydrologique

- James Bennett, Durga Lal Shrestha and David Robertson for the Commonwealth Scientific and Industrial Research Organisation (CSIRO, Australia) verification case study for ephemeral streams

- Valeria Koli (Russian Federal Service for Hydrometeorology and Environmental Monitoring (ROSHYDROMET), Russian Federation)

- Yashar Falamarzi (Climate Research Institute (CRI), Atmospheric Science and Meteorological Research Centre (ASMERC), Islamic Republic of Iran)

- Jihoon Park (Nakdong River Flood Control Office (NRFCO), Republic of Korea)

Finally, I would like to extend my deep appreciation to the members of the WMO Standing Committee on Data Processing for Applied Earth System Modelling and Prediction (SC-ESMP), the WMO Expert Team on Operational Hydrological Prediction Systems (ET-OHPS) and the WMO Research Board for the final review of this publication, and to the WMO Secretariat (Hwirin Kim, Giacomo Teruggi and Rokhaya Ba) for their support in developing this activity.

Ian Lisk

President, Commission for Weather, Climate, Hydrological, Marine and Related Environmental Services and Applications (SERCOM)

CHAPTER 1. INTRODUCTION

Accurate hydrological forecasts are essential to safeguarding lives and properties, improving the quality of life, enhancing the economy and protecting the environment. To that end, all countries invest in and operate some form of hydrological forecasting infrastructure and provide forecast products and services derived therefrom. Hydrological forecasts serve a wide range of users. At one end, there are emergency managers who may be faced with the possible evacuation of large communities within a short amount of time and water managers who must operate complex systems of reservoirs and pipelines to supply water while minimizing risks from flooding or drought. At the other end, there are individual residents and motorists who may have to take immediate action to avoid fast-rising waters or flooded roadways.

With continuing human-made changes to land cover and the environment as well as climate change, hydrological forecasting is an increasingly important yet challenging enterprise (NRC, 2001). It is widely recognized that hydrological verification has a larger role to play in improving the accuracy and quality of hydrological forecasts and allowing for more effective use of the forecast information in users' decision-making (Welles et al., 2007). For many practitioners of hydrological forecasting, however, hydrological verification remains largely outside of routine hydrological operations. The goal of this publication is to help promote the practice of hydrological verification by providing the community with a practical guide for initiating and conducting verification.

Verification refers to the process of comparing the forecast of interest with the verifying observation to assess the accuracy and quality of the forecast. Verifying observation refers to the observation with which one may ascertain the occurrence or non-occurrence of the forecast event or how good the forecast is. Accuracy refers to the representative quantitative measures of forecast error. The measures may vary depending on the type of the forecast (for example, single-valued or probabilistic, categorical or continuous). Quality refers to the collective characteristics, or attributes, of the forecast that translate into skill useful for decision-making. Skill refers to statistical measures of forecast accuracy relative to some reference such as a climatological forecast.

To the producers of hydrological forecasts, verification serves two main purposes. The first is to cost-effectively improve the accuracy and quality of the forecast by objectively guiding systematic improvement of the end-to-end forecast process and the forecast systems used therein. The forecast process may include, but is not limited to, observing, modelling, calibration, data assimilation (DA), analysis, prediction and postprocessing. The second is to communicate the accuracy, quality, skill and information content of the forecast based on past performance to the users of the forecast information so that they may use the real-time forecast more effectively for user- or application-specific decision-making. To the users of the forecast, verification information allows calibration, be it explicit or implicit, of the user's decision support and decision-making systems and processes, which increases the utility and value of the forecast information.

Hydrological verification is rooted in weather forecasting and verification of weather forecasts. As with any forecast verification, the science and practice of hydrological verification draw heavily from statistics, probability theory and information theory. Whereas many publications and other resources exist on verification science, the literature is limited on practical guidance for initiating hydrological verification in the real world. Despite the long-standing, widespread and often very sophisticated use of statistics, applied probability and stochastic processes in hydrology and water resources engineering, many practising hydrologists and civil engineers may find the language of verification too unfamiliar to readily explore its content for possible adoption and practice.

This publication is an effort to help bridge such gaps and facilitate the practise of hydrological forecast verification in the real world. Specifically, it is aimed at providing both the producers and the users of hydrological forecasts with a practical introductory guide for hydrological

forecast verification, including the science, logistics, and freely available tools and resources. Key concepts are presented to elicit intuitive understanding in the context of hydrological applications in the real world. Wide-ranging real-world examples are provided for hands-on verification with step-by-step instructions. With these guidelines, readers will be able to replicate the results using the data, tools and resources identified herein and start practising hydrological verification using their own.

Hydrological forecasting is subject to large uncertainties in the meteorological forecasts used as input and in the hydrological, hydraulic and water management models used to describe the movement and storage of water. Hydrological forecasting also covers multiple scales; the space–time scale over which the predictive skill and uncertainty must be captured and represented with physical-dynamical and statistical consistency is often very large, ranging from local to continental scale in space and from minutes to decades in time. With increasing urbanization and interconnectedness in commerce and people's lives, there is also an ever-growing and universal need for increased spatio-temporal specificity and lead time in hydrological forecasts. Higher-resolution hydrological forecasts with longer lead times, however, can only be provided with increased uncertainty. Operational hydrological forecasts nevertheless must serve a wide range of applications that require varying levels of forecast information and a broad spectrum of users with varying levels of risk perception and tolerance.

In working towards addressing the collective challenges and needs expressed above, the ensemble approach has emerged in recent years as the most practical one for operational hydrological forecasting (Wells, 2017). Ensemble forecasting is well suited for processing weather and climate forecasts, which are inherently uncertain, and provides an estimate of predictive uncertainty which conveys a measure of confidence in the forecast, thereby allowing for user-specific risk-based decision-making. With the use of longer-range ensemble forcing forecasts, hydrological ensemble forecasting may extend lead time (Kim et al., 2018), and multimodel ensemble forecasting may improve forecast accuracy, as premised by information theory (Georgakakos et al., 2004).

Single-valued (for example, deterministic) forecasts may be considered a special case of ensemble forecast in which only a single member exists in some representative form such as mean or median. Therefore, the mechanics of verifying ensemble forecasts encompass those of verifying single-valued forecasts. One may hence use ensemble verification tools for verification of single-valued forecasts if necessary. For example, one may calculate the root mean squared error (RMSE) of single-valued forecasts by calculating the RMSE of ensemble mean forecasts. One may also calculate the mean absolute error (MAE) of single-valued forecasts by calculating the mean continuous ranked probability score (CRPS) (see Chapter 4 for more information about the measures mentioned above). This does not mean, however, that one may interpret the resulting verification statistics for single-valued forecasts as if they are for ensemble forecasts.

Most probabilistic verification measures apply to all types of probabilistic forecasts. Most probabilistic forecasts are in the form of ensemble forecasts with ensemble members representing outcomes that are equally likely to occur, or probability distribution forecasts with empirical or parametric probability distributions describing the range and relative likelihood of possible outcomes. Depending on the metric, the mechanics of performing verification and interpreting the results differ by varying extents between ensemble forecasts and probability distribution forecasts. Given the widespread use and increasing adoption of hydrological ensemble forecasting, this publication deals primarily with ensemble forecasts for probabilistic verification and makes a distinction whenever necessary between the two to avoid potential confusion.

1.1 Scope and approach

These introductory guidelines are focused on verification of streamflow, or discharge, at specific locations and deals with verification of input forcing forecasts from numerical weather

prediction (NWP) only to a very limited extent. To many users, verification of river stage, rather than discharge, is often of greater and more immediate interest for flood preparedness and emergency evacuation purposes. Therefore, operational streamflow verification should, in general, include both stage and discharge forecasts. Logistically, verification of stage forecast is effectively the same as verification of discharge forecast, and to verify stage forecast one may follow the steps described in this document for verification of discharge.

Stage–discharge relationships, which are described by rating curves, are heavily modulated by channel geometry and hydraulic properties at and near the gauging station. Hence, stage forecast is not very reflective of the predictive skill of the forecast system and predictability of the hydrometeorological and hydrological processes over the drainage area. In addition, verification of stage forecast does not, in general, allow comparative assessment of forecast quality, including uncertainty propagation, throughout the river basin in reflection of the movement and storage of water from upstream to downstream locations through channels, reservoirs and outlet structures. If one is interested in using verification to aid systematic improvement of forecast quality for the entire hydrological system of interest, it is hence necessary to verify streamflow forecasts. Stage verification, however, represents location-specific hydraulic translation of streamflow forecast quality via rating curves or hydraulic models.

The above scope means that this publication does not address verification of flash floods or any other predictands of importance in hydrological forecasting other than streamflow, such as precipitation, temperature, snow water equivalent and other variables. The intent is that the information, resources and tools shared herein will also be useful, directly or indirectly, for the verification of other variables and phenomena of interest for hydrological forecasting.

In addition, this publication does not address the assessment of utility (Benjamin and Cornell, 1970) or value of streamflow forecasts. Readers interested in the assessment of the socioeconomic value of hydrological services derived from hydrological forecasts are referred to *Valuing Weather and Climate* (WMO-No. 1153) and Laugesen et al. (2023). Assessing the socioeconomic value of streamflow forecasts requires relevant decision support systems and cost–loss models (Murphy, 1969; Zhu et al., 2002; Laio and Tamea, 2007). The former translates the forecast information to application-specific decisions. The latter translates decisions to user-specific and/or societal value. The cost–loss models for streamflow forecasts are often highly nonlinear and require consideration of a wide range of spatio-temporal scales. For example, the cost associated with not taking any precautionary action for a high-flow event is usually far greater if the observed peak flow exceeds the critical flood stage than if it crests just below that stage. For the operation of a multipurpose reservoir for a heavy rainfall event, the cost of precautionary action and the loss from inaction may conflict with the actions needed to optimize flood control and water supply, and there may also be conflict between the interests of residents upstream of the reservoir and those downstream. Often, the largest socioeconomic loss from a flood event is caused by widespread inundation, particularly in urban areas. If inundation forecasts are not available from hydrodynamical models (Noh et al., 2019), it is necessary to develop inundation maps based on location-specific streamflow forecasts (NWS, 2012). As may be seen from the discussion above, models for decisions, cost and loss can be rather complex and may require multidisciplinary expertise and resources. Regardless of the complexity, the ensemble approach is well-suited for assessment of the value of streamflow forecasts in that one may propagate the ensemble members through the chain of models and assess the socioeconomic value expressed in the output.

The users of hydrological forecast information are very diverse, and so are their risk perception and tolerance. Advanced users may base their risk tolerance on modelling and analysis results from decision support tools and risk-based quantification of costs and benefits, whereas others may only qualitatively update their risk perception. Similarly, the predictability of streamflow varies greatly among different hydroclimatological regions and often from one location to another even within the same region, and so does the predictive skill among different forecast systems and processes. Given this picture, this publication does not attempt to produce a fixed set of universal guidelines. Instead, it provides general guidance with key points and offers

case studies and examples in which the guidance is put into practice. Key points are provided at the end of each chapter and summarized collectively in Chapter 8. In this way, readers may follow the general guidance for their verification task and arrive at a core set of verification metrics, scores and diagrams that address their verification objectives.

With the wide-ranging verification examples provided, users with limited data and resources will still be able to initiate streamflow verification, albeit at a limited level and scope (the verification tools are freely available). The data and resource limitations should not, however, deter readers from gaining experience with hydrological verification and taking steps towards wider practices. For this reason, this publication places particular emphasis on gaining a fundamental understanding of forecast verification beyond model validation so that the users will be able to make the most of the available data, identify possible weak links in their forecast systems and processes, and develop and plan for cost-effective pathways towards improving forecast products and services with the aid of verification. An important element distinguishing verification from validation is the provision and communication of verification information to the users of the forecast products and services so that they may make better decisions. Such verification information also aids forecasters to better understand their forecasts and better explain them to their customers, which will in turn help build stronger cases for the necessary data and resources.

1.2 Organization

The organization of this publication as an introductory guide reflects its ambitious goal, that is, guide readers unfamiliar with verification step by step from gaining an understanding of the fundamental concepts in the context of hydrological forecasting to being able to conduct hydrological verification using one of the examples in Appendix A as a template. This publication assumes that the reader is trained in hydrology and water resources engineering and has basic knowledge of applied probability and statistics represented in basic flood frequency analysis (Linsley et al., 1982). Given that topics such as estimation and modelling of tails of probability distributions are an integral part of statistical hydrology and water systems design, one might argue that hydrologists and water resources engineers are particularly well prepared to wade into verification.

Chapter 2 addresses the general question of why one may want to practice verification, including what spurred the science and practice of verification in weather forecasting, how verification differs from other forms of evaluation, and how verification may lead to improving the accuracy and quality of hydrological forecasts and therefore their utility and value.

Chapter 3 describes the qualities that make forecasts most useful for decision-making. Firm understanding of such “attributes of forecast quality” is essential to utilizing all available verification metrics, scores, and diagrams effectively and making sense of the verification results in the context of the real world. To aid readers in intuitively understanding the fundamental concepts, Chapter 3 heavily utilizes graphical illustrations in the context of flood forecasting. This is a tutorial chapter intended for those who are not familiar with verification and may be skipped on the first reading of this document.

Chapter 4 describes and explains the verification metrics, scores and diagrams commonly used in water and weather forecasting. Strengths, limitations, potential pitfalls and possible adaptations are also described in the context of hydrological forecasting and applications.

Chapter 5provides guidance on the steps necessary before embarking on verification as an organized activity and describes the logistical issues to consider.

Chapter 6 provides guidance on displaying verification results and creating visual verification information for various users.

Chapter 7 presents six wide-ranging case studies of hydrological verification as listed in Table 1.

Table 1. List of case studies in hydrological verification presented in Chapter 7

Case number	*Case name*	*Tools used*
1	Verification of single-valued streamflow forecast with uncertainty decomposition	EVS[a]
2	Comparative verification of multiple single-valued streamflow forecasts	EVS
3	Verification of ensemble streamflow forecast for headwater and downstream locations	EVS
4	Verification of skill in ensemble streamflow forecast for water supply	EVS
5	Diagnostic verification in (near) real time: Système de Prévision Hydrologique (Government of Québec, Canada); Performance Analysis Tool (Australian Bureau of Meteorology); and Rijkswaterstaat Operational Systems (Kingdom of the Netherlands)	
6	Comparative verification of ensemble forecasts for ephemeral streams	MATLAB

[a] Ensemble Verification System

Chapter 8 provides a summary of key points.

Appendix A presents seven hands-on examples of streamflow verification using various verification tools as listed in Table 2.

Table 2. List of hands-on examples of streamflow verification provided in Appendix A

Example number	*Example name*	*Tools used*
1	Verification of single-valued streamflow forecast with uncertainty decomposition	EVS
2	Verification of ensemble streamflow forecast with aggregation of multiple forecast points	EVS
3	Verification of ensemble streamflow forecast with skill score and confidence interval calculations	EVS
4	Computational example 1	EVS
5	Computational example 2	R verification package
6	Computational example 3	Python verif library

Example number	*Example name*	*Tools used*
7	Comparative verification of accuracy, reliability and sharpness of two ensemble streamflow forecasting systems in a stream with zero flow	MATLAB

Information about the Ensemble Verification System (EVS) (Brown et al., 2010), R verification package and Python verif library may be found in computational examples 1, 2 and 3, respectively.

Appendix B provides the mathematical definitions and expressions for the distributions-oriented approach (Murphy and Winkler, 1987) with graphical illustrations, and those for the various statistical moments referred to throughout this document. This appendix is included for completeness and reference and may be skipped on the first reading of this publication.

CHAPTER 2. WHY VERIFY?

All forecasts, be they deterministic or probabilistic, are subject to errors whose magnitude and characteristics are part of the forecast information. Hence, if the quality of the forecast is unknown or is not communicated to the forecasters and the end users, the forecast can only be described as incomplete (NRC, 2001). Verification supports "completing the forecast" by doing the following: (1) providing the forecast users with up-to-date forecast performance information for better decision-making (Brier and Allen, 1951; Murphy and Winkler, 1987; Murphy, 1993; Welles et al., 2007); (2) providing the operational forecasters with identification and objective assessment of the strengths and weaknesses of the forecasts and forecast systems and the uncertainties therein; and (3) providing the forecast developers with objective comparative assessment of the newly developed forecast versus the existing one and objective guidance for forecast system enhancements (Demargne et al., 2009).

2.1 Early work in forecast verification

As operational weather forecasting started in the United States of America and in Europe during 1850–1870, the meteorological community began to question the quality of weather forecasts and started to discuss the concepts and methods for forecast verification (Burton, 1986; Murphy, 1996). One of the earliest papers on the subject was by Finley (1884) and describes the accuracy of an experimental tornado forecast system in the United States. In this paper, Finley, a pioneer in forecasting severe storms, reported an overall accuracy of 96.6% for the tornado forecast system. This figure was based on the percentage of correct tornado and no-tornado forecasts, accounting only for the correct positive and correct negative (that is, no tornado) forecasts.

Table 3. Pooled results of Finley's experimental tornado forecasting programme

	Observations		
Forecasts	*Tornado*	*No tornado*	*Total*
Tornado	28	72	100
No tornado	23	2 680	2 703
Total	51	2 752	2 803

Source: Murphy (1996)

At first glance, Finley's accuracy metric may appear to represent very skillful forecasts. However, it is heavily influenced by the very large number of correct negatives (2 680 out of 2 803 forecasts) without accounting for the large number of false alarms (72) and misses (23) relative to the number of tornadoes observed (only 51 tornadoes were observed out of 2 803 forecasts). When false alarms and misses were considered, only 55% of the tornadoes were correctly predicted, and 72% of the forecasts for tornadoes were false alarms. Moreover, a naïve forecast that states that there will never be a tornado would result in an accuracy of 98.2% by Finley's scoring (for the data and verification scores, see Table 3 and https://www.cawcr.gov.au/projects/verification/Finley/Finley_Tornadoes.html). Finley's paper (1884) spurred the publication in 1884–1893 of several papers on the deficiencies of the percentage correct score in tornado forecast verification, issues related to forecast verification in general, and the development of alternative verification methods and measures, some of which are still widely used today (Murphy, 1996).

As illustrated by the so-called "Finley affair" described above, many concepts, methods and practices in forecast verification were spearheaded by the meteorological community, which recognized very early the importance of verification in improving weather forecasting. These verification methods are now increasingly used in other disciplines, including hydrology (Troin et al., 2021; Anctil and Ramos, 2019).

2.2 Modes of hydrological verification

Forecast verification is the process of evaluating forecast quality as measured by the degree of correspondence between the forecast and the reference (usually the verifying observation) or by the skill of the forecast relative to some reference forecast (for example, a naïve forecast based on climatology or persistence). In the context of hydrological forecasting, verification may take on three different but complementary modes of operation:

(i) Diagnostic verification may be performed whenever data availability or increase in sample size warrants generating or updating the necessary or desired verification information.

(ii) Real-time verification may be performed in near real time to aid predictive assessment of the quality of a live forecast before the outcome is observed.

(iii) Event verification may be performed after a significant event to assess the quality of the forecasts issued and the performance of the forecasting system used for the specific event as part of post-analysis.

Increasingly often, diagnostic verification is performed in (near) real-time, that is, as soon as or very soon after the verifying observation becomes available. Such "operational" verification provides near-immediate feedback to forecasters, as elaborated upon in section 5.2.2. Case 5 in Chapter 7 provides three different real-world applications of such verification.

Conceptually, one may consider real-time verification a form of diagnostic verification but conditioned on the current environmental conditions. Such conditioning identifies the past forecasts or reforecasts (if generated retrospectively) that share similar meteorological and hydrological conditions to the current one. For example, if heavy rainfall is expected from a tropical storm over wet soil, past forecasts issued under similar conditions, or historical analogues, are extracted (for example, from a relational database) and verified in near real time. Such conditional verification necessarily assumes adequate sample size so that the resulting verification information enhances predictive skill. If the number of analogues is too small relative to the number of attributes considered in the environmental conditions, the resulting conditional verification information is likely to suffer from overfitting and hence lack predictive skill.

If the logistics allow, one may avoid real-time verification by performing diagnostic verification conditionally on pre-identified sets of important meteorological and hydrological attributes in advance. In practice, such conditioning events can only be defined rather coarsely, as sample size decreases very quickly with each additional conditioning attribute. In addition, one must accept the risk that the conditioning might turn out to be incorrect, in which case the use of the resulting verification information may potentially be counterproductive. Real-time verification should only be performed by operational forecasters who already have the full knowledge of the verification information without the event-specific conditioning and know how to synthesize such additional verification information.

The scope of event verification often extends beyond the assessment of forecast quality to include that of the timeliness and appropriateness of warnings. In addition, evaluations of the performance of the service delivery components of a hydrological forecasting system may also be included, such as an assessment of the degree to which the forecasting and forecast-informed decision procedures were followed and whether "perfect warnings" (that is, the set of

warnings that should have been issued as per procedures) would have captured the severity and impact of the event to a satisfactory degree. In event verification, forecast quality is assessed using the forecasts produced immediately before and during the event and all observations available at the time of the analysis. Therefore, sample size is small, and the forecast–observation pairs only reflect the single event. One must hence recognize that the resulting verification information does not represent the overall quality of the forecasting system.

Well-planned verification should identify and address specific questions about forecast quality to effectively inform the decisions of the forecast users. Below are several groups of such questions:

- How suitable are the forecasts for a given application? Are the forecasts sufficiently unbiased for the decisions to be made? For example, probabilistic forecasts should not systematically under- or over-prescribe the probability of flooding. Is the forecast better than naïve forecasts? Multiple metrics may be used to assess various aspects of forecast quality. For example, a probabilistic flood forecast may be evaluated in terms of the accuracy of the forecast probability of exceeding the flood level and the ability of the forecast to capture the flood peak timing.

- What are the strengths and weaknesses of the forecasts? For which cases does the forecast system perform well or poorly? For example, is the forecast quality different for high- versus low-flow conditions, specific seasons (winter versus summer) or the state of the atmosphere or the hydrological system (for example, wet versus dry antecedent soil moisture)?

- What are the important sources of error in the forecast? Is the forecast impacted more by errors in the forcing inputs or errors in the initial conditions (ICs)? Multiple forecast streams may be set up to isolate the impact of different sources of error in the forecast chain (for example, forcing inputs, ICs, model parameters, model structure) as well as their interactions.

- How are new science and technology improving the forecasts? Do the forecasts from the new forecast system improve the verification results over those from the current system? The new system may reflect new calibration parameters, a new preprocessing or postprocessing technique, or a new source of observation. Verification helps objectively assess the impact on forecast accuracy and quality of any new development of the forecasting system and track the forecast improvement over time.

- What should be done to improve the forecasts? Verification should offer guidance on prioritizing the forecast system development and enhancements.

Given the diversity of the verification questions and that of the potential users of the verification information, different levels of verification are needed to enable informed decision-making (Jolliffe and Stephenson, 2012). For the above, one must clearly identify the objectives of the verification task and the specifics of the verification information sought so that the appropriate forecast–observation datasets and verification methods may be determined (Brier and Allen, 1951). As noted in section 1.1, these guidelines are focused on the evaluation of forecast quality, which is a necessary step for assessing the forecast value for specific user decisions. As part of verification planning, it is hence a good practice to consider the range of decisions that the users may make to increase the information content of the verification results, even if one does not explicitly assess the forecast value in the verification task at hand.

2.3 Types of forecasts and observations

Verification relies on the comparison of the forecast event with an observation, or a high-quality estimate, of its outcome. The forecast may be single-valued or probabilistic. Depending on the type of forecast, the choice of verification methods and metrics will vary. Whereas single-valued forecasts provide no information about the uncertainty or confidence in the single predicted value (Wilks, 2011), probabilistic forecasts describe the range and likelihood of possible outcomes. An ensemble forecast may be easily converted to a probabilistic forecast. For example, when 25% of the ensemble members rise above the flood level at some future time, there is a 25% chance of flooding at that time according to the ensemble forecast. Ensemble forecasts are sometimes collapsed into single-valued forecasts by retaining only the ensemble mean or median. Any measure of the uncertainty captured by the ensemble members is then lost, and the quality of the ensemble forecast can only be described partially. One can attach uncertainty bounds to a single-valued forecast based, for example, on flow-dependent error distributions derived from a large sample of past forecasts and verifying observations. Such practices are in fact common in hydrological forecasting (Regonda et al., 2013). Once such uncertainty information is attached, however, one no longer has a single-valued forecast but a probabilistic forecast or, if one samples plausible equally-likely realizations from it, an ensemble forecast.

Forecasts, whether single-valued or probabilistic, may describe binary (for example, flood and no flood), categorical (for example, major flood, minor flood and no flood), or continuous (for example, streamflow) predictands. Binary, or dichotomous, forecasts may be derived from the ensembles, for example, by defining the occurrence of an event as the forecast probability exceeding some threshold. Similarly, an ensemble forecast may also be converted to a categorical forecast by considering multiple thresholds. For example, with two thresholds – for minor and major flooding – one may prescribe the probabilities of major and minor flooding from an ensemble forecast. Multiple categories may also be defined based on quantiles. One may use the observed streamflow values associated with exceedance probabilities of, for example, 10%, 30%, 70% and 90% as the thresholds for different categories. The proportion of the ensemble members in each category or the cumulative proportion up to and including each category could then define a probabilistic forecast. A multicategory forecast may also be treated as a set of binary forecasts by considering each threshold separately.

The verifying observation may be discrete, with only a limited set of possible outcomes (for example, the two outcomes of flooding and no flooding), or continuous, with an infinite number of possible outcomes. Given a threshold, a continuous observation may be expressed as an indicator variable which takes on the value of 1 if the observation exceeds the threshold, and 0 otherwise. For a categorical probabilistic forecast, the indicator variable for the observation may, for example, be defined as 1 for all categories greater than or equal to the category in which the observation falls, and 0 for all other categories. Table 4 summarizes the classification of hydrological forecasts according to type and specificity, with examples.

Table 4. Classification of hydrological forecasts according to type and specificity, with examples

Type	*Specificity*	*Examples*
Single-valued (deterministic)	Binary	Flooding or no flooding
	Categorical	Minor/major flooding for two-category flood forecast
	Continuous	Streamflow, ensemble mean forecast of streamflow, water level or stage, volume
Probabilistic	Binary	Probability of flooding

Type	*Specificity*	*Examples*
	Categorical	Probabilities of minor and of major flooding, probabilities of streamflow exceeding the 10th, 30th, 70th and 90th percentiles of observed flow
	Continuous	Probability distribution of streamflow, ensemble streamflow forecast, ensemble water level (or stage) forecast

Most verification studies, including those in this publication, assume that observation errors are negligibly small compared to forecast errors. Observations are subject to both random and systematic errors, such as biases and representativeness errors in measurement, reporting errors, conversion errors, and analysis errors which occur when the observational data are analyzed or remapped to match the scale of the forecast. Most streamflow "observations" are in fact estimates obtained via rating curves rather than "measurements". Therefore, streamflow observations are subject to significant uncertainties, particularly for large flows. Harmel et al. (2006) report streamflow observation errors for small watersheds of 42%, 19%, 10%, 6% and 3% for the worst case, typical maximum, typical average, typical minimum and best-case scenarios, respectively. Di Baldassarre and Montanari (2009) report that the overall error affecting river discharge observations ranges from 6.2% to 42.8%, at the 95% confidence level, with an average value of 25.6%. Bowler (2008) reports that observation errors reduce the apparent skill of the forecast system, and that their effect is typically largest at short lead times when forecast errors are smallest. When verifying seasonal streamflow forecasts, streamflow observations may be naturalized to take out the effects of dams or diversions. In such cases, the resulting naturalized streamflow "observations" would necessarily be subject to larger errors. Even if one does not explicitly account for observation errors in the verification process, it is a good practice to consider how the verification results and any conclusions therefrom might potentially be impacted by significant observation errors.

2.4 Sources and types of error and roles of calibration, validation and verification

Forecast errors are collective realizations of the total uncertainty associated with the predictand (that is, the variable being predicted). There are two large sources of uncertainty in hydrological forecasting: forcing input uncertainty and hydrological uncertainty (Krzysztofowicz, 1999; Seo et al., 2006). Forcing input uncertainty includes all uncertainties associated with the hydrometeorological forecasts used as input, such as quantitative precipitation and temperature forecasts from the numerical weather and climate prediction models. Hydrological uncertainty includes all uncertainties associated with the hydrological, hydraulic, reservoir and any other water models used in hydrological forecasting under no input uncertainty (that is, under clairvoyant hydrometeorological forecasts). Hydrological uncertainty typically includes uncertainties in the model structures, parameters, and initial conditions (ICs), and those associated with human control of movement and storage of water (known or unknown). Input uncertainty increases with lead time as hydrometeorological variables become increasingly less predictable further into the future.

Structural uncertainty arises due to the deficiencies in hydrological, hydraulic and other water models, such as lack of model physics (for example, infiltration) and poor model dynamics (for example, routing). Since no model is ever perfect, structural uncertainty always exists regardless of the choice of models. Parametric uncertainty comprises the uncertainty in the tunable model parameters and the uncertainty in the various geographical information system (GIS) layers and other location-specific physiographic attributes used to prescribe the fixed boundary conditions (BCs) in the hydrological, hydraulic and other water models. The IC uncertainty is associated with the model states valid at the prediction time where forward integration of the models begins. Because the model states are commonly not observed directly in operational hydrological forecasting, particularly at the catchment scale (soil moisture being the prime example), the ICs are generally subject to large uncertainties. For

this reason, they are kept as up to date as possible in operational forecasting by re-running the models over the very recent past using all available observations, including late-arriving ones.

One of the most cost-effective (and hence routinely practised) ways to reduce hydrological uncertainty in hydrological modelling and prediction is to reduce parametric uncertainty via calibration (that is, by adjusting the tunable model parameters to maximize the agreement between the model simulation and the verifying observation under some criteria) (Duan et al., 2006). Calibration typically addresses parametric uncertainty only. Hence, the performance metrics primarily used in calibration, such as the Nash–Sutcliffe efficiency (Nash and Sutcliffe, 1970) and the Kling–Gupta efficiency (KGE) (Gupta et al., 2009), do not, in general, measure the set of forecast attributes that are necessary to assess predictive hydrological uncertainty. If desired, one may use indices such as KGE in forecast verification. There are, however, potential pitfalls in using scoring rules that are not strictly proper (see section 4.2.2), particularly when the forecasts are unfamiliar (see Appendix B). For this reason, care is necessary in adopting performance metrics used in calibration for verification.

In real-time forecasting, some form of state updating is also routinely practised to reduce the IC uncertainty using manual or automatic DA of real-time observations, those of discharge being the most important. The positive impact of state updating is generally large for short lead times but, depending on the memory of the hydrological system, wears off relatively quickly as the lead time increases. Hydrological uncertainty may also include anthropogenic uncertainties if there are unknown or poorly known human-made changes in the movement and storage of water. For example, forecast of outflow from a dam will have much larger uncertainty if the reservoir release schedule or the operating rule is unknown or incorrectly known. Figure 1 provides a qualitative depiction of the input and hydrological uncertainties versus lead time.

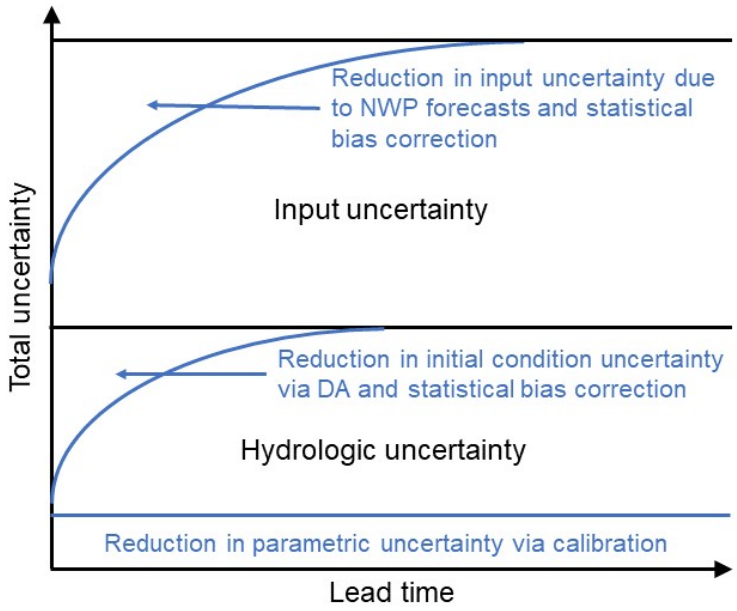


Figure 1. Qualitative depiction of input and hydrological uncertainties versus lead time

From the perspective of hydrological forecasting, one may consider input and hydrological uncertainties as aleatory and epistemic, respectively. The former is caused largely by the random or chaotic variations in the Earth system, whereas the latter arises more from the lack of knowledge of the phenomena, processes, parameters or states than randomness or extreme sensitivity to ICs or BCs. Epistemic uncertainty most often manifests itself as systematic errors, or biases, whereas aleatory uncertainty is usually realized as both random and

systematic errors. If random errors have little statistical structure, they tend to cancel out when aggregated over space or time. Systematic errors, on the other hand, tend not to cancel out, resulting in biases. Assessing both input and hydrological uncertainties is an important aspect of hydrological verification. Based on the relative magnitude of the two, the forecasting agency may identify the more cost-effective areas of improvement. Case 1 in Chapter 7 describes a simple example of decomposing total uncertainty into input and hydrological uncertainties, and Example 1 in Appendix A provides a hands-on example of Case 1 using the EVS.

Type I and type II errors are most often associated with statistical hypothesis testing. In verification, type I, or false positive, and type II, or false negative, errors refer to incorrect predictions of an event, such as flooding, that did not and did occur, respectively. The two types of errors are competing attributes of a forecast in that, unless the accuracy of the forecast is improved, they cannot in general be reduced simultaneously. The above point is illustrated in Figure 2 using a toy example made with synthetically generated data. The figure shows scatterplots of two sets of short-range river stage forecasts, forecast A (black) and forecast B (red), versus the common verifying observation. The two forecasts are based on the same time series model, but different DA techniques for state updating are used for real-time error correction. The DA technique for forecast A considers type I error only, whereas that for forecast B considers both type I and type II errors (Shen et al., 2022a). In the figure, the vertical and horizontal lines at 1.8 m represent a fictitious bankfull stage above which flooding begins to occur. If the forecast and the verifying observation fall within the false positive or negative region demarcated by the above lines as shown in the figure, the forecast has type I or II error, respectively. Note that forecast A shows no type I (false positive) error but large type II (false negative) error and fails to score even a single hit. Forecast B shows smaller type II (false negative) error but larger type I (false positive) error compared to forecast A. The example illustrates that, depending on the cost or the negative impact associated with type II versus type I error, one may arrive at very different conclusions about the utility of a forecast or the relative utility of competing forecasts. Hence, assessment of type I and II errors is an important part of verification.

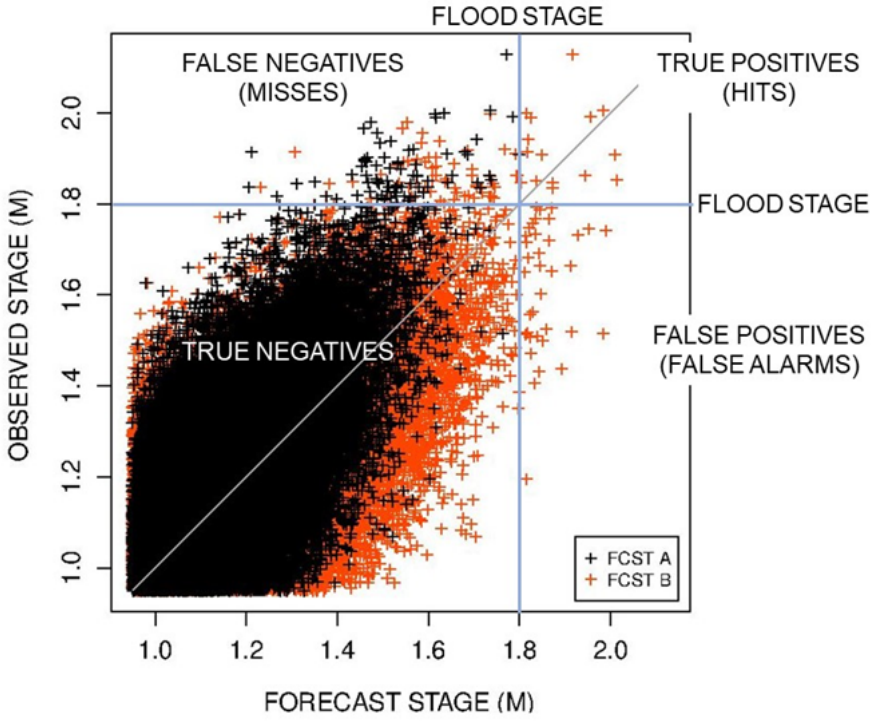


Figure 2. Illustration of type I (false positive) and type II (false negative) errors

For flood forecasting, accurate prediction of time to peak and the timing and magnitude of the rising limb and peaks of the hydrograph is of particular importance. Timing errors arise from various sources, in addition to the errors in the precipitation forecast, as described below. After a prolonged inter-storm period, the soil moisture states of rainfall-runoff models are subject to increased uncertainty. Hence, significant errors are more likely in forecasting the timing of the incipient rise of the hydrograph from significant rainfall after a long dry spell. Unless corrected in a timely manner via state updating in real time, such timing errors generally lead to large errors in streamflow forecasts, particularly in the rising limb (Shen et al., 2022b). If unit hydrograph (UHG) is used to route surface runoff for a catchment that is too large for the UHG assumptions to hold consistently, rainfall concentrated in one part of the catchment is likely to result in timing errors in streamflow forecasts at the catchment outlet even with perfect rainfall forecasts. When routing flood waves through a river reach using observed upstream BCs, timing errors may result from inaccurate fixed BCs and fluid-dynamical parameters in the routing or hydrodynamical model (Kim et al., 2021). Because the BC and the parameters vary with the physiographic properties of the channel (for example, in-channel versus flood plain), timing errors vary with the magnitude of the flow and the physiography of the flow paths. The above examples suggest that the verification information for time-to-peak forecasts, by itself, is not likely to be very useful for decision support or diagnosing the performance of the forecast system, and that time-dependent characterization and quantification of the errors in streamflow forecasts is necessary to contextualize timing errors in the reference frame of the hydrograph response. Such time-dependent verification requires taking the time dimension into account and hence poses a higher-dimensional problem of identification, estimation and inference compared with typical streamflow verification. Verification of streamflow forecasts for different stages of the hydrograph response and that of time-to-peak forecasts is an area of further research and presents new opportunities for advancing the science and practice of hydrological verification (Liu et al., 2012).

The terms verification and validation are used somewhat differently in different disciplines (Oreskes et al., 1994). In geoscience, verification refers to objective comparison of forecasts and observations to establish accuracy of the forecast. Validation, on the other hand, refers to the substantiation that the process and methodology are working correctly and producing the intended results (Wilson, 2017). In machine learning, independent validation is routinely practised to assess performance, and cross-validation is widely used for this purpose. Cross-validation partitions the data into two subsets and uses one subset for training, or calibration, of the model and the other for validation. Typically, multiple rounds, or folds, of cross-validation are performed using different partitions to reduce sampling variability, after which all validation results are combined into one for calculation of performance measures. In hydrological or meteorological verification, there is no training or calibration involved, and in the context of machine learning, only a single round of independent validation is performed, in most cases using all available forecasts within the spatio-temporal domain of interest.

Bootstrapping is another resampling method widely used in machine learning (Efron, 1979). In verification, bootstrapping is often used to estimate confidence intervals by randomly sampling historical forecasts with replacement, performing verification, and repeating the above steps many times over to generate an ensemble of the same verification statistics. While computationally expensive, bootstrapping is relatively straightforward and makes no distributional assumptions, hence it is a powerful tool for assessing sampling uncertainty. Examples 3 and 7 in Appendix A provide hands-on applications of bootstrapping for estimation of confidence intervals in ensemble streamflow verification using the EVS and MATLAB, respectively.

For large-to-extreme events, sample size may be too small to produce statistically significant verification results that are location-specific. In such cases, some type of regionalization or trading of space for time is likely to be necessary, similarly to the approaches used in precipitation and streamflow frequency analyses (Perica et al., 2018; Riggs, 1973). Most forecasts have limited dynamic ranges when compared to the verifying observations, which often leads to under- and over-prediction of very large and small events, respectively. The above situation may result in type II error, or false negatives, which are particularly important

for prediction of large-to-extreme events (see Chapter 3 for a conceptual illustration). Type II error is often masked by sampling uncertainty and may not readily reveal itself in verification statistics. Hence, it is very important to visualize the raw forecasts in some form and scrutinize them against the verifying observations as part of the verification process (see Chapter 6). This practice is also important for assessing the practical significance of the differences in the verification statistics for competing forecasts.

2.5 Verification of hydrometeorological forecasts

Hydrological forecasts are usually forced by meteorological forecasts of precipitation, temperature, and possibly other variables such as potential evapotranspiration. The skill in the hydrological forecasts is hence usually bounded by the skill in the meteorological forecasts (see Figure 1). Verification of the forcing input forecasts is an important aspect of hydrological verification as it allows assessment of the relative importance of input versus hydrological uncertainties and the impact of the quality of the input forecast on that of the hydrological forecast. Weather forecasting centres usually produce verification statistics along with their operational weather forecasts. Such verification information, however, is generally of very limited utility in hydrological verification as explained below (see also discussions in Pappenberger et al. (2008) and Anderson et al. (2019)).

Meteorological forecasts are gridded and spatially continuous, and are usually verified regionally or over large areas. Streamflow forecasts, on the other hand, are verified at specific locations where observations are available, typically at the basin outlets or points of interest along river reaches. In addition, the domain and resolution of the meteorological forecasts are not specific to the catchment areas and generally are not comparable to the spatio-temporal resolution of the hydrological models used to produce hydrological forecasts. Even if the spatio-temporal resolution of the input forecasts matches that of the hydrological model (if a distributed model or models are used), the verification information over a large area is not representative of the specific catchments of interest, particularly for precipitation and temperature in complex terrain with strong orographic influences (Harris et al., 2001; Brussolo et al., 2008). Note that a precipitation forecast that is accurate at a regional scale to a meteorologist can very easily be a complete miss to a hydrologist or a water resources engineer, particularly for small basins if precipitation falls outside of the boundary of the catchment of interest.

It is therefore necessary to verify the input forecasts used in operational hydrological forecasting for the impact assessment or the uncertainty decomposition described above. Such verification often reveals catchment- and lead time-specific biases in precipitation and temperature forecasts that are also scale- and terrain-sensitive (Pappenberger and Buizza, 2009; Imhoff et al., 2020). Most hydrological ensemble forecast systems include a component to remove or reduce such biases, the Meteorological Ensemble Forecast Processor (MEFP) (Schaake et al., 2007; Wu et al., 2011) for the Hydrologic Ensemble Forecast Service of the United States National Weather Service (NWS) (Demargne et al., 2014) being an example.

2.6 Key points

- The purpose of hydrological verification is to increase the value and utility of hydrological forecast products and services by supporting objective and systematic improvement of forecast quality and the decisions of the users of the forecast information.

- Hydrological verification broadly utilizes the theory and practices developed by the meteorological community, which was early to recognize the value of verification in improving weather forecasting.

- Hydrological forecasts are subject to input and hydrological uncertainties. The former are associated with errors in the hydrometeorological forecasts used as input to hydrological models. The latter are associated with errors in the rest of the hydrological forecasting process. Verification supports uncertainty decomposition to guide cost-effective improvement of forecast input, systems and processes.

- Type I, or false positive, and type II, or false negative, errors are competing attributes of a forecast. Verification informs the trade-off between the two types of errors and supports decision-specific assessment of the utility of a forecast and the relative utility of competing forecasts.

- Movement and storage of water is heavily modulated by the physiography of the individual catchments, river basins, channels and water bodies. Hence, unlike weather forecasts, hydrological forecasts should be verified as location-specifically as possible to the extent data availability allows.

- Prediction of large-to-extreme events is very often the most important service of operational hydrological forecasting. Such events occur infrequently, and hence the sample size tends to be small. To increase sample size, some form of trading of space for time or regionalization is usually necessary at the expense of location specificity.

CHAPTER 3. ATTRIBUTES OF FORECAST QUALITY

To assess the quality of a forecast, the forecast is compared, or verified, against an observation or a high-quality estimate of the outcome. In some cases, the verifying observation may not be an actual observation or derived from one but a model output such as a hydrological simulation that makes it possible to isolate errors from a particular source or sources. Forecasts of higher quality display a stronger correspondence with the observations, as may be assessed qualitatively in scatterplots of single-valued forecasts versus verifying observations. Such scatterplots are empirical representations of the joint probability distribution between the forecast and the verifying observation from which the relevant conditional and marginal distributions may also be obtained (see Appendix B for the mathematical relationships). Verification theory describes and summarizes forecast quality in the reference frame of this joint distribution.

3.1 Introduction

To describe the joint distribution above, it is necessary to describe, or model, the forecast and the verifying observation as random variables. In this section, we use a simple toy example to illustrate how such modelling gives rise to an unknown joint distribution between the forecast and the verifying observation, what verification is concerned with regarding this distribution and what assumptions are commonly made in practical verification. This stand-alone section is for tutorial purposes only and is intended for those who may be familiar with statistics but not necessarily with basic probability theory.

First, let us denote the discrete random variables representing the forecast and the observation of some uncertain variable of interest as X and Y, and the experimental values, or outcomes, that X and Y may take on as x and y, respectively. Let us now consider forecasting rain or no rain in a neighbouring town a day ahead. One may make a nearly effortless forecast by looking at the sky for cloudiness. If the cloud cover does or does not exceed some threshold (referred to as cloudy or not cloudy for brevity), one forecasts rain or no rain, respectively. For verification, a similarly effortless observation may be made by looking out the window the next day. If one does or does not see rain on the ground from one's house, one considers that it did or did not rain in the neighbouring town as well, respectively.

Cloudiness alone is generally a very poor model for forecasting rain or no rain a day ahead and hence one may expect large forecast errors. Due to the spatial variability of rainfall, local and regional hydroclimatological variations, and possible lack of representativeness of the location of one's house, looking out the window may not be a very good model for observing rain or no rain in a neighbouring town and hence may lead to significant observation errors.

In the illustrative example above, both the forecast and observation errors are largely random in nature, and hence the forecast and the observation can only be described as random variables. For verification, the look-at-the-sky model represents the (categorical) forecast, X, whose possible outcomes are cloudy and not cloudy. Similarly, the look-out-the-window model represents the (categorical) observation, Y, whose possible outcomes are rain and no rain. Note that the possible outcomes for each model are mutually exclusive (that is, only one of the two can occur at any given time) and collectively exhaustive (that is, nothing else but the two can ever occur). One may assign 1 and 0 (or some other numbers or symbols) to the outcomes of cloudy and not cloudy, respectively, in which case the experimental value, x, of the random variable, X, is 1 or 0. Similarly, one may assign 1 and 0 to the outcomes of rain and no rain, respectively, in which case the experimental value, y, of the random variable, Y, is 1 or 0. One may surmise from the above that there likely exists a relationship, albeit a weak one, between the two random variables (X and Y) which may be described by their joint, or bivariate, probability distribution. In the context of the example above, verification is concerned with identifying the most important descriptors, or attributes, of the unknown joint distribution between X and Y, estimating the statistics for the attributes, and making

inferences about the joint distribution. The above identification, estimation and inference are made based on the outcomes or the experimental values, *x* and *y*, of the forecasting and observing models, *X* and *Y*, from many repeated "experiments" using the look-at-the-sky and look-out-the-window models, respectively. Table 6 provides an example of a tabulation of the results of many such experiments, whereas a scatterplot of single-valued forecasts versus verifying observations is an example of a visualization of the results of many such experiments for continuous random variables (see, for example, Figure 2). Of course, in operational water and weather forecasting in the real world, the forecast and observation models are far more sophisticated, but the concepts are the same.

An important assumption behind the above approach is that each forecast–observation pair is independent (of all other pairs) and identically distributed (IID) (and hence stationary). The IID assumption entails that, in theory, one only needs to describe a single joint relationship between the two random variables representing the forecast and the observation. Whereas the outcomes of experiments such as tossing a coin many times (that is, a Bernoulli process (Drake, 1967)) may indeed be considered IID, the assumption is not likely to hold all inclusively for variables such as precipitation and streamflow. Therefore, it is often necessary to resample and stratify the forecast–observation pairs so that the subsampled pairs may be considered to share a common joint distribution or condition the pairs to make inferences about certain parts of the distribution.

The (unknown) joint relationship between forecast and observation is fully described by an infinite number of statistical moments such as mean, variance and skewness (see Appendix B). To capture the essence of the relationship, multiple aspects or attributes of forecast quality are necessary. Hence, verification requires the use of different measures and scores (Murphy, 1993, 1997). The choice of metrics to provide sufficient information about the forecast quality may vary depending on the forecast type (single-valued or probabilistic), the specificity of the forecast (dichotomous, categorical or continuous), the forecast application and the user needs, in addition to the predictability of the variables being forecast and the predictive skill of the forecast systems and processes. It suffices to say that verification measures and scores should be chosen to give the user meaningful information on which the user can make informed decisions (Jolliffe and Stephenson, 2012).

3.2 Organization of this chapter

Most verification metrics, scores and diagrams are complementary, but many overlap with one another by varying degrees in information content. For effective and efficient verification, it is necessary to identify the important forecast attributes for the verification task at hand and narrow the large array of metrics, scores and diagrams down to a core combination that sheds the most light on the important, as well as differentiating (good or bad), qualities of the subject forecast. Ideally, the metrics of choice should be largely independent of one another in information content to avoid tangential analysis (that is, more is not necessarily better). For the above, it is necessary to gain firm understanding of and familiarity with several fundamental concepts and the "menu" of available metrics, scores and diagrams. Chapter 3 and Chapter 4 address the above.

The rest of this chapter describes the various aspects of forecast quality. To aid those who are not familiar with probabilistic forecasting, the attributes are described first in the context of single-valued forecasts and then extended to probabilistic forecasts. For single-valued forecasts, scatterplots showing the forecasts and their verifying observations provide an intuitive depiction of forecast quality; in general, the tighter the scatter around the one-to-one line, the higher the forecast quality. In addition, one may easily relate the widely used verification measures such as the mean error (ME), RMSE and correlation with the general geometry of the scatter in most cases.

For probabilistic forecasts, the measures and diagrams used to assess forecast quality (see Chapter 4) are more complex. To utilize them effectively, it is necessary first to acquaint

oneself with several attributes such as reliability, resolution, uncertainty, type II conditional bias, discrimination and sharpness. Though equally applicable except for sharpness, the above attributes are not often used in the verification of single-valued forecasts, and hence many may find them unfamiliar. To aid intuitive understanding, this chapter provides pictorial explanations of the attributes first using idealized single-valued forecasts and then using simplified ensemble forecasts in physical space, rather than in probability space, in emulation of the scatterplots with which hydrologists and water resources engineers are familiar.

3.3 Forecast attributes in the context of single-valued forecasts

This section describes selected attributes that are frequently used in verification. They are first described in the context of single-valued forecasts to avoid potential confusion that may arise from navigating between physical space and probability space.

3.3.1 Bias

Bias refers to systematic error in the forecast relative to the verifying observation. The most widely used form is mean bias, or first-order bias, defined as the difference between the average of the forecasts and that of the verifying observations. Climatological forecasts, by definition, have no mean bias. Streamflow forecasts that tend to over- or under-forecast have positive or negative mean bias, respectively.

One may similarly define second-order bias as the difference in standard deviation (or variance) between the forecast and the verifying observation. If the variability of the forecast is smaller than that of the observation due, for example, to the limited dynamic range of the forecast system, the standard deviation of the forecast is likely to be biased low, that is, the forecast will have a negative second-order bias. If the forecast suffers from large random errors, it may have a positive second-order bias.

Higher-order moments and biases are often not considered in the verification of single-valued forecasts for two main reasons. The first is that the commonly used moment-based verification metrics such as the mean squared error (MSE) are only of second order. The second is that estimation of higher-order moments requires an increasingly larger sample size which is often not available in practice. However, precipitation and streamflow, arguably the two most important variables in operational hydrological forecasting, generally have skewed (that is, asymmetric) distributions (Bras and Rodriguez-Iturbe, 1984). Skewness is an important indicator not only of the shape of the tail of the distribution and but also of possible heteroscedasticity (that is, nonuniformity in variability), both of which are particularly important for forecasting and verification of large-to-extreme events. Significant heteroscedasticity is an indication that the IID assumption may not be reasonable. Therefore, examination of skewness and heteroscedasticity often sheds significant additional light on the assessment of magnitude-dependent predictability and predictive skill in hydrological verification (Pagano and Garen, 2005). The Breusch–Pagan test (Breusch and Pagan, 1979), developed originally for linear regression, is very useful for testing heteroscedasticity and is available in various statistical packages, including in multiple R packages.

3.3.2 Correlation

Correlation refers to the strength of statistical association between the forecast and the verifying observation. The relationship may be linear or nonlinear, positive or negative. The two most widely used measures for association are the Pearson correlation, or simply "correlation", which measures linear correlation, and the Spearman's rank correlation, or rank correlation, which measures ordinal association. Rank correlation is often used to assess the strength of the monotonic relationship between forecasts and observations.

3.3.3 Accuracy

Accuracy refers to the level of agreement between the forecast and the verifying observation. The most widely used measure of accuracy for single-valued forecasts is the MSE (or RMSE) which reflects mean bias, second-order bias and correlation (see Equation 61 in Appendix B). Accurate forecasts form a tight scatter with verifying observations around the one-to-one line. Such forecasts necessarily have small mean bias, small second-order bias and high correlation.

3.3.4 Skill

Skill refers to the relative accuracy of the forecast compared to some reference forecast or benchmark of choice. If the subject forecast is more accurate than the reference forecast, the former is said to have skill, or to be skillful. A skill score calculates fractional improvement in accuracy by the subject forecast over the reference forecast using the measure of accuracy of choice.

The general definition of a skill score for the given metric and reference forecast (such as climatology or the forecast from a baseline forecasting system) is given by:

$$Skill\ Score\ =\ \frac{Metric_{fcst}\ -\ Metric_{ref}}{Metric_{perfect}\ -\ Metric_{ref}} \tag{1}$$

where $Metric_{fcst}$, $Metric_{ref}$ and $Metric_{perfect}$ are the values of the chosen accuracy metric for the forecast being evaluated, the reference forecast, and the perfect forecast, respectively. If the score of the perfect forecast is equal to 0 (which is the case for most metrics, including the MSE), the skill score is given by:

$$Skill\ Score\ =\ \frac{Metric_{ref}\ -\ Metric_{fcst}}{Metric_{ref}} \tag{2}$$

As an example, if the measure of accuracy chosen is the MSE, the MSE skill score (MSESS) of the subject forecast with respect to the reference forecast is given by:

$$MSESS = \frac{MSE_{ref} - MSE_{fcst}}{MSE_{ref}} = 1 - \frac{MSE_{fcst}}{MSE_{ref}} \tag{3}$$

where MSE_{fcst} and MSE_{ref} are the MSEs of the subject and reference forecasts, respectively.

The skill score of a perfect forecast is unity. A skill score of zero means that the forecast is no better than the reference forecast under the chosen measure of accuracy. A negative skill score indicates that the subject forecast is less accurate than the reference forecast. The choice of the reference forecast depends on the purpose of the verification. Usually, the reference forecast is a "naïve" forecast, such as (observed) climatology, persistence (defined as the most recent observation) or random chance. The benchmark could be a forecast produced from a baseline forecasting system if the aim is to assess improvements in the forecast system. The Nash–Sutcliffe efficiency (Nash and Sutcliffe, 1970), which is widely used in hydrology for model calibration, is a MSE skill score with observed climatological mean as reference.

3.3.5 Decomposition of mean squared error

Murphy and Winkler (1987) introduced a general verification framework referred to as the distributions-oriented approach in which the forecast and observation are treated as random variables and the forecast–observation pairs are assumed to be IID (see section 3.1). In this approach, forecast quality is assessed by factorizing the joint probability distribution of the

forecast and the verifying observation into conditional and marginal distributions. Appendix B provides the pertinent mathematical details.

A key element of the above approach is the decomposition of the MSE in two different ways: the calibration–resolution (CR) decomposition and the likelihood–base rate (LBR) decomposition (Murphy and Winkler, 1987). The former decomposes the MSE into reliability (REL) (also referred to as type I conditional bias), resolution (RES) and uncertainty (UNC), and the latter decomposes the MSE into type II conditional bias (T2B), discrimination (DIS) and sharpness (SHA):

$$MSE = REL - RES + UNC \tag{4}$$

$$MSE = T2B - DIS + SHA \tag{5}$$

The mathematical expressions of all terms in Equations 4 and 5 may be found in Appendix B, and the expressions for their statistical estimators may be found in Chapter 4. Equations 4 and 5 are referred to hereafter as the CR and LBR decomposition equations, respectively. Mathematically, the LBR decomposition equation is identical to the CR decomposition equation, except that the variables for the forecast and the verifying observation are interchanged. For example, UNC in the CR decomposition equation represents the variability (measured by variance) of the observation. Hence, SHA in the LBR decomposition equation does the same for the forecast. The difference between the CR and LBR decomposition equations is that the former assesses forecast accuracy, as measured by the MSE, in reference to the forecast (that is, conditioned on the forecast event), whereas the latter does the same but in reference to the verifying observation (that is, conditioned on the observed event). Hence, one may consider the CR decomposition a forecast-centric attribution of forecast accuracy, whereas the LBR decomposition may be viewed as an observation-centric attribution. Though verification and prediction are different, one may think of CR and LBR decompositions as being analogous to forward (that is, regular) and reverse linear regression, respectively.

The above two decompositions result from factorizing the same joint probability distribution of the forecast and the verifying observation in two different ways. Hence, the terms in the CR and LBR decomposition equations may be more easily understood with the help of idealized scatterplots representing empirical probability density functions (PDFs) or probability mass functions (PMFs) (see Appendix B for the distinction), as shown in Figure 3 and Figure 4. In each panel of the figures, the parallelogram represents the outermost boundary of a uniformly spread and infinitely dense scatter of the single-valued forecast–observation pairs. Their empirical joint PDF, or histogram, is hence a parallelepiped. In the figures, the blue areas reduce the MSE (hence, the larger they are, the more desirable), whereas the red areas increase the MSE (hence, the smaller they are, the more desirable) in accordance with the CR and LBR decomposition equations above. Note that if the forecasts are perfect and hence the parallelograms are compressed into the diagonal line, the blue areas are at maximum and the red areas vanish.

3.3.6 Reliability

Reliability measures the mean bias between the forecast event and the verifying observations for a specific forecast event. The forecasts are said to be reliable if the average of the verifying observations for a specific forecast event, that is, the conditional mean of the verifying observations for a forecast event, is the same as the forecast event itself for all forecast events. For example, single-valued streamflow forecasts of 500 m^3/s are reliable, or conditionally unbiased in the type I error sense, if the average verifying observed flow is 500 m^3/s. Climatological forecasts, albeit not very accurate, are reliable because the average verifying observation is always the same as the forecast itself.

Figure 3 shows an idealized example of perfectly reliable single-valued forecasts versus the verifying observations. The blue dots represent the averages of all verifying observations for

the specific forecast events selected. In Figure 3(a), because the above averages lie on the midpoints of the imaginary vertical lines within the parallelogram, the blue dots line up on the dashed 45° line, resulting in REL = 0 and hence no visible red area. Note in Figure 3(a) that the forecasts in this idealized example are never larger or smaller than the largest and the smallest conditional mean of the verifying observations, respectively. The above geometry suppresses false positives (that is, type I error) and hence type I conditional bias. Recall that, in flood forecasting, a false positive, or false alarm, occurs when flooding is forecast but does not occur (see section 4.2). Figure 3(a) illustrates that reliability measures type I conditional bias, and that a departure of the blue dots from the diagonal in either direction indicates that the forecasts are less than reliable.

3.3.7 Resolution

Resolution measures the forecast's ability to differentiate between different observed outcomes. For example, if the forecasts have good resolution, the average verifying observed flow for forecasts of 500 m^3/s is significantly different from that for forecasts of 200 m^3/s. Resolution is measured in Figure 3(a) by the average, over all forecast events, of the square of the vertical distance of a blue dot from the mean of all observations indicated by the dotted horizontal line. Hence, the blue triangles in Figure 3(a) represent the area contributing to RES. Pictorially, resolution is reflected in the overall slope of the blue dots. If the slope is zero, the observed outcome is the same regardless of the forecast, and hence RES is zero (this is akin to the slope in regular linear regression). If the slope is non-zero, the larger the slope is, the larger the resolution is. Note, however, that, for the same MSE (that is, given the same accuracy), increasing resolution (increasing the magnitude of RES in the CR decomposition equation) can only be achieved at the expense of deteriorating type I conditional bias (that is, increasing REL in the CR decomposition equation).

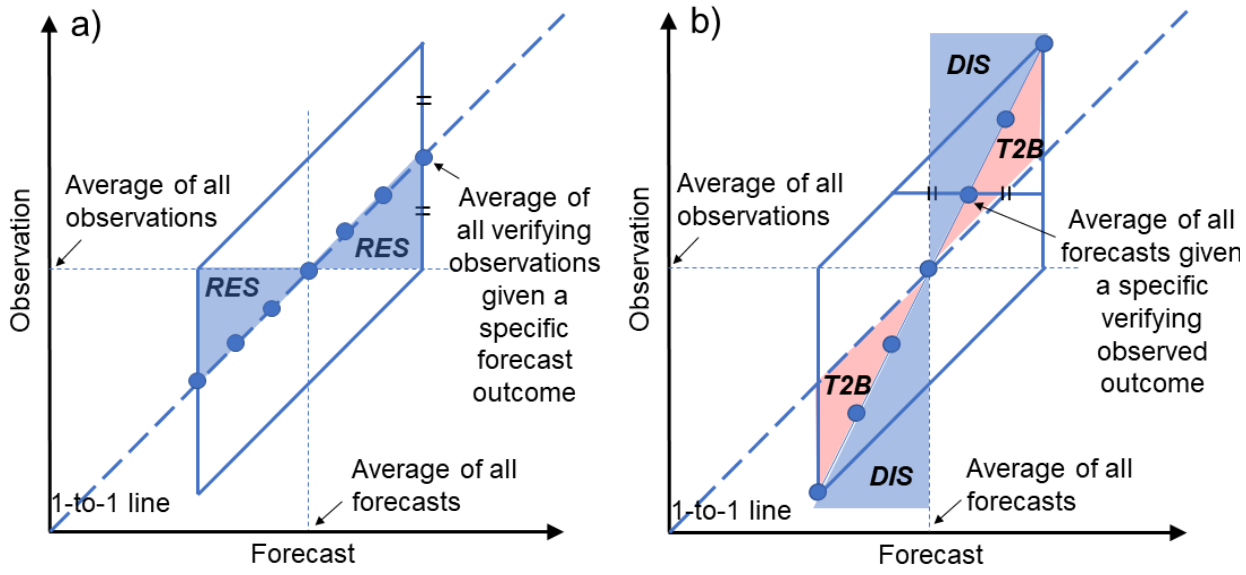


Figure 3. Idealized scatterplot of forecasts with no type I conditional bias (REL = 0) versus verifying observations under (a) CR decomposition (into REL and RES) and (b) LBR decomposition (into T2B and DIS). The blue and red areas contribute positively and negatively to the MSE, respectively.

3.3.8 Uncertainty

Uncertainty is a measure of the variability of the predictand, and is given by the variance of the observation. For streamflow and precipitation, which are generally skewed and heteroscedastic, UNC may easily stretch from a few to several orders of magnitude. Often, the relative importance of different forecast attributes varies significantly with the magnitude of variability in UNC. For streamflow verification, assessment of UNC with respect to various physiographic attributes provides very useful guidance on possible stratification, data pooling

or conditioning of the verification dataset and the overall verification strategy. Such attributes may include season, flow regime, flow magnitude, hydrograph response, whether the locations are in the headwaters or downstream, and others. Flow duration curves (Searcy, 1959), which have been used in hydrology and water resources engineering for many decades, are useful for such purposes.

3.3.9 Type II conditional bias

Type II conditional bias measures the mean bias in the forecast for specific observed events. For example, if the single-valued flow forecasts for the common verifying observation of 500 m^3/s have a mean of 700 m^3/s, the forecasts have a positive type II conditional bias of 200 (= 700 – 500) m^3/s. Figure 3(b) shows the same scatterplot shown in Figure 3(a) but with the LBR decomposition. A blue dot in Figure 3(b) represents the average of all forecasts associated with a specific observed event (that is, the conditional mean of the forecasts for an observed event). Because the average of all forecasts along a horizontal line within the parallelogram is the midpoint of that horizontal line segment, the blue dots line up along the longer diagonal of the parallelogram. If the above conditional mean is the same as the observed event across all observed events, the forecasts have no type II conditional bias. The red triangles in Figure 3(b) hence represent the area contributing to type II conditional bias over all observed events.

Figure 3 shows that the idealized forecasts, while perfectly reliable, are conditionally biased in the type II error sense. The parallelogram in this idealized example is bounded by two vertical sides. Hence, the dynamic range of the forecast is always smaller than that of the verifying observation, leading to false negatives, or misses, and hence type II conditional bias. A false negative occurs in flood forecasting when flooding is not forecast but does occur (see section 4.2).

3.3.10 Discrimination

Discrimination measures the degree by which the forecasts are different for different observed events. If the average streamflow forecast for high-flow events tends to be different from that for low-flow events, the forecasts are discriminatory. Climatological forecasts are not discriminatory, since the average forecast is the same regardless of the observed event. In the LBR decomposition equation, DIS is defined as the difference squared, averaged over all observed events, between the average of all forecasts associated with a specific observed event and the average of all forecasts. The difference above is represented in Figure 3(b) by the distance between a blue dot and the dotted vertical line representing the average of all forecasts. Hence, the blue triangles in Figure 3(b) represent the area contributing to DIS. Analogously to RES, DIS is reflected in the overall slope of the blue dots. If the slope is infinitely large (or if it is zero in reverse regression in the context of regression), the forecasts are the same regardless of the observed event and hence DIS is zero.

Analogously to Figure 3, one may also consider the MSE decompositions under no type II conditional bias. Figure 4(a) and (b) show the LBR and CR decompositions, respectively, in a second idealized scatterplot. Note that, whereas the parallelogram in Figure 3 has two vertical sides, that in Figure 4 has two horizontal sides. Although the forecasts in Figure 4 have no type II conditional bias (that is, T2B = 0, as the blue dots line up on the diagonal in Figure 4(a)), they are conditionally biased in the type I error sense (that is, they are less than reliable, with a positive REL corresponding to the red areas in Figure 4(b)). The inherent conflict in reducing REL versus reducing T2B without reducing MSE reflects the fact that improving reliability and reducing type II conditional bias is a zero-sum game, that is, zero T2B in Figure 4 can only be achieved by issuing a larger number of forecasts that are outside of the range of the observed events, which increases false positives, or false alarms, and hence reduces reliability.

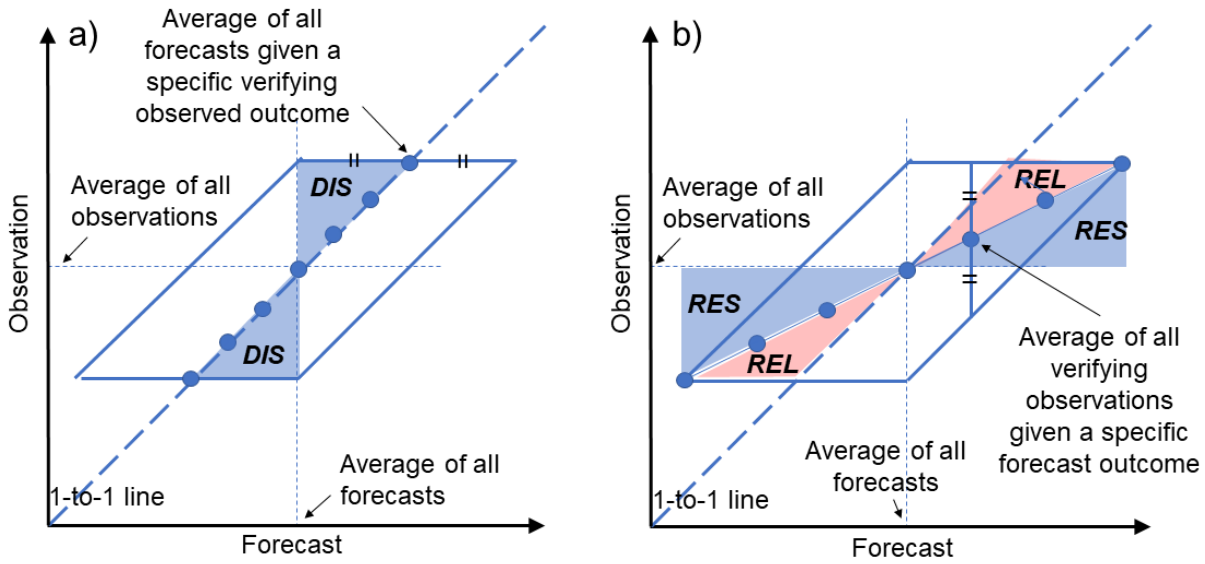


Figure 4. Idealized scatterplot of forecasts with no type II conditional bias versus verifying observations under (a) LBR decomposition (into T2B and DIS components) and (b) CR decomposition (into REL and RES components). The blue and red areas contribute positively and negatively to the MSE, respectively.

3.3.11 Sharpness

The SHA component is a measure of the variability of the forecast and is given by its variance. Single-valued forecasts with very limited dynamic range have very small sample variance and hence small SHA. Note in the LBR decomposition equation that, given the same MSE, SHA can only be reduced at the expense of deteriorating type II conditional bias or discrimination. Conversely, given the same MSE, type II conditional bias or discrimination can only be improved by increasing the variability of the forecast. Figures 3 and 4 illustrate the above point in that the elimination of type II conditional bias in Figure 4(a) is achieved by increasing the dynamic range of the forecast (hence larger SHA) in comparison with Figure 3(b).

Though SHA is described above under the heading of sharpness, the common definition of sharpness for probabilistic forecasts is different from the definition of SHA. This is a potential source of confusion and is explained in section 4.2.2 with the aid of the statistical estimators for the components of the LBR decomposition of the Brier score (BS).

3.3.12 Illustrative example

With the definitions of the attributes associated with the MSE decomposition for single-valued forecasts on hand, we now return to the toy example of Figure 2 and comparatively assess the attributes of forecasts A and B. Figure 5(a) is analogous to Figures 3(a) and 4(b) and may be used to assess reliability and resolution. The solid cyan circles and triangles in Figure 5(a) represent sample means of the verifying observations conditional on forecasts A and B falling within each subrange indicated by the gray vertical bars, respectively. In the legend, E[OBS|FCST A] denotes the mean observed stage given that forecast A falls in a specific subrange. Other similar notations are analogously defined.

Figure 5(a) indicates that, within its dynamic range, forecast A is very reliable (very small REL) and has good resolution (large RES), whereas forecast B is less reliable (larger REL) and has reduced resolution (smaller RES). One may also surmise that forecast A has better resolution within its dynamic range than forecast B from the larger slope of the scatter associated with forecast A. It is important to note in Figure 5(a) that sample conditional means for forecast A could not be calculated for the two largest subranges due to the limited dynamic range of forecast A. Hence, REL and RES for forecast A, by themselves, provide little information about

the quality of the forecast for large-to-extreme events whose magnitude may exceed the dynamic range of the forecast.

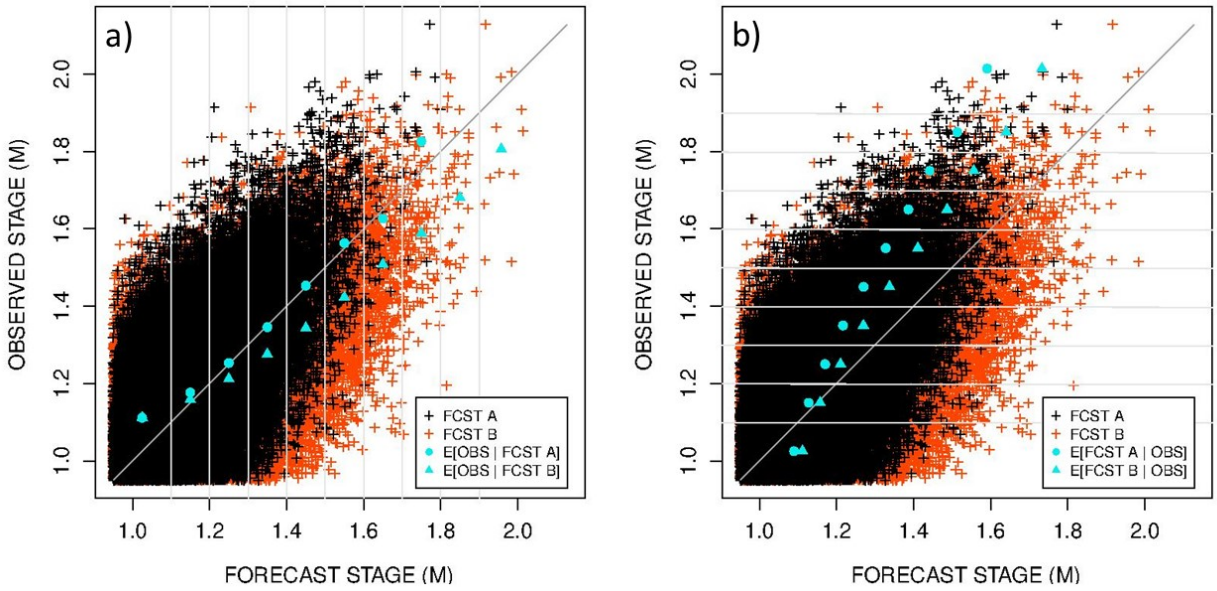


Figure 5. Scatterplots of forecasts A and B versus verifying observations. The solid markers denote the averages of all verifying observations/forecasts within the respective vertical/horizontal strips. The distance between a solid marker and the diagonal represents type I/II conditional bias in the forecast specific to the corresponding subrange of the forecast/verifying observation.

Figure 5(b) is analogous to Figure 3(b) and Figure 4(a) and may be used to assess type II conditional bias and discrimination. The solid cyan circles and triangles in Figure 5(b) represent sample means of forecasts A and B conditional on the verifying observation falling within each subrange indicated by the gray horizontal bars, respectively. Figure 5(b) indicates that forecast A has large type II conditional bias (large T2B) and relatively modest discrimination (modest DIS), whereas forecast B has smaller type II conditional bias (smaller T2B) and better discrimination (larger DIS) than forecast A. One may also surmise that forecast B has better discrimination than forecast A from the smaller slope of the scatter associated with forecast B. Forecasts A and B have sample standard deviation of 0.12 and 0.15 (m), respectively. Hence, forecast B has larger SHA than forecast A in agreement with the observation in the scatterplots that forecast B has a larger dynamic range than forecast A.

With the key attributes reviewed above, it is instructive to consider an example of how verification may help improve hydrological forecasting in a changing (that is, nonstationary) world. A case in point is the balanced reduction of type I and type II errors, an age-old challenge in estimation and prediction in general (Lieberman and Cunningham, 2009). With urbanization, land cover changes and climate change, accurate hydrological forecasting of large-to-extreme events is an increasingly important yet challenging endeavour. For such events, the negative consequences of a false negative (that is, failing to see the wolf) are often significantly larger than that of a false positive (that is, crying wolf when there is none). Hence, objective and coherent assessment of the trade-offs among different forecast attributes and their possible variations across different flow regimes is an increasingly important element of operational hydrological forecasting for which hydrological verification is essential.

3.4 Forecast attributes in the context of probabilistic forecasts

This section illustrates the various aspects of forecast quality for probabilistic verification using hypothetical ensemble forecasts. To aid intuitive understanding, the ensemble forecasts are represented in box-and-whisker plots as shown in Figure 6 in emulation of scatter plots for single-valued forecasts. An ensemble forecast is a probability distribution defined empirically

by its ensemble members. The box and whiskers for each ensemble forecast provides a simplified representation of this probability distribution. In Figure 6, each box contains the 20th to 80th percentiles of the ensemble members, with the median indicated by the bar within the box. The whiskers extend from the smallest to the largest ensemble member.

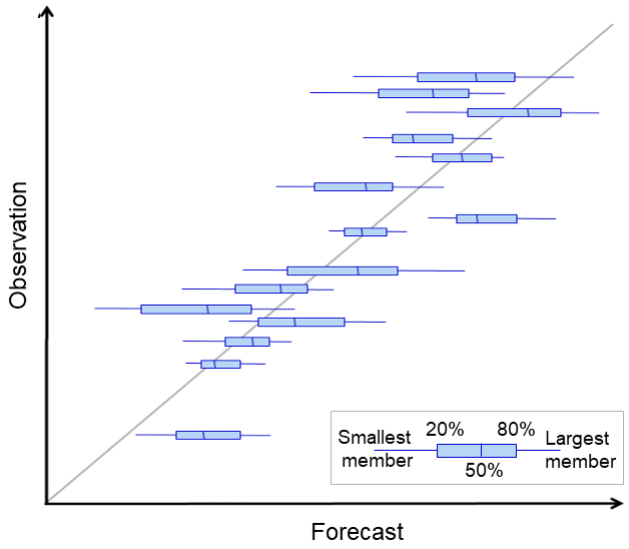


Figure 6. Scatterplot-like representation of hypothetical ensemble forecasts and verifying observations. Each ensemble forecast is represented by a box and whiskers. The solid line shows the 1:1 relationship between the ensemble forecasts and the observations.

In the following subsections, the most widely used attributes of probabilistic forecasts are described (Wilks, 2011; Jolliffe and Stephenson, 2012; Murphy, 1997; Bradley et al., 2019). In the conceptual illustrations, the hypothetical ensemble forecasts are for streamflow. In some examples, the ensemble streamflow forecasts are used to forecast the probability of streamflow exceeding some threshold (such as a flood level). In these cases, the exceedance probability is given by the fraction of the ensemble members above the threshold. The verifying observation for a probability forecast is its indicator variable representation, that is, 1 if the observed flow exceeds the threshold, and 0 otherwise. The fraction of the observations exceeding the threshold in the verification dataset gives the observed frequency of flooding.

3.4.1 Bias

For probabilistic forecasts, bias refers to systematic error in forecast probability and hence relates to questions such as: Are the probabilistic flood forecasts consistently over- or under-prescribing the probability of flooding? If so, there is a systematic bias in the probabilistic forecasts. If the forecast probability of streamflow exceeding some threshold is consistently lower or higher than the fraction of the times when the verifying observed flow exceeds the threshold, the probability forecast has a negative (or low) or positive (or high) bias, that is, a tendency to under- or over-forecast probability, respectively.

In verification of probabilistic forecasts, reliability (that is, type I conditional bias) and type II conditional bias are routinely assessed directly or indirectly. If the probabilistic forecast is conditionally unbiased for all conditioning events, it is unconditionally unbiased, or simply unbiased. Hence, assessing conditional biases also assesses (unconditional) unbiasedness in probabilistic verification. For ensemble forecast, unconditional bias in distribution may be assessed without assessing conditional bias via the rank histogram, also known as the Talagrand diagram (see Chapter 4). The rank histogram does not consider the joint relationship between the forecast and the verifying observation but only checks if the ensemble members and the verifying observation are mutually independent realizations of the

same probability distribution (Stephenson and Jolliffe, 2003), that is, if the verifying observation is statistically just another ensemble member of the forecast distribution.

3.4.2 Correlation

For verification of probabilistic forecasts, the strength of association between the forecast probability and the categorized verifying observation is reflected in reliability, resolution, type II conditional bias and discrimination following the BS decomposition (see section 4.3.2 and 0). For this reason, correlation is not separately considered in probabilistic verification.

3.4.3 Accuracy

For probabilistic forecasts, accuracy refers to the level of agreement between the probability forecasts and the verifying observations. The verifying observation for a probability forecast is its indicator variable representation: 1 if the forecast event is observed and 0 otherwise. The most widely used measure of accuracy for probabilistic forecasts of binary events is the Brier score (BS), which is a MSE of probability forecasts. The most widely used measure of accuracy over the entire range of a probabilistic forecast is the continuous ranked probability score (CRPS). The CRPS is the integration over the range of the forecast of the difference between the forecasted exceedance probability and the step function representation in probability space of the verifying observation squared (that is, integration of the BS over the range of the forecast). The BS and CRPS are described in detail in Chapter 4. The smaller the CRPS, the more accurate the probabilistic forecast is. Figure 7(a) illustrates highly accurate ensemble forecasts which agree closely with the observed outcomes (and hence a small mean CRPS). In contrast, the ensemble forecasts in Figure 7(b) exhibit significantly larger forecast errors, which indicates lower accuracy (and hence a significantly larger mean CRPS).

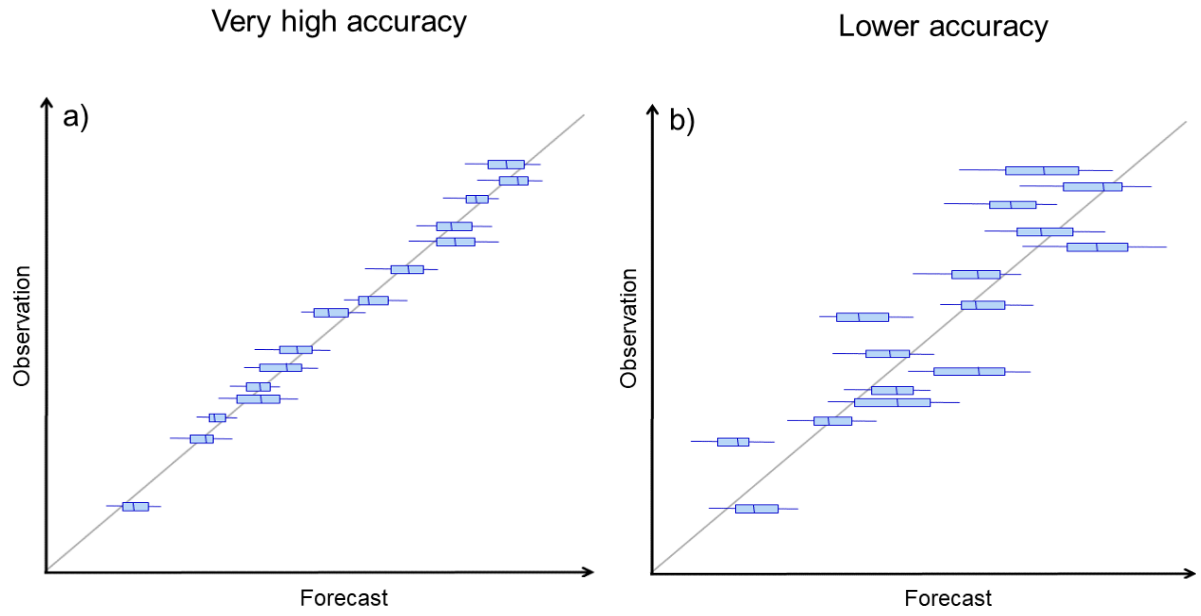


Figure 7. Illustrative examples of (a) highly accurate ensemble forecasts and (b) ensemble forecasts with lower accuracy with the one-to-one line overlaid

3.4.4 Skill

As with single-valued forecast, skill for probabilistic forecast describes the accuracy of the subject forecast relative to a reference probabilistic forecast or benchmark of choice. The choice of the reference depends on a number of factors, including the temporal scale of interest. For example, short-term forecasts for the next few hours must be more skilful than persistence, which assumes that the current condition, such as the most recently observed flow or water level, will persist into the future. Climatology should be explicitly defined when

used as a benchmark for assessing skill. Unconditional climatology should be estimated from a long record of observations, ideally for a longer period than the sample climatology from the verification dataset. However, seasonal climatology (climatology of all observed outcomes over many years stratified according to season), calendar day-specific climatology (climatology for the same day of the year across all historical years) or climatology from some other time scale may provide a more useful baseline. Pappenberger et al. (2015) discuss how to choose a benchmark forecast to evaluate hydrological ensemble forecasts and how to avoid benchmarks that are too simplistic (see references therein for examples of benchmarks based on climatology, persistence and simplified models).

Skill scores for probabilistic forecasts are defined in the same way as those for single-valued forecasts (see Equation 1). The BS and mean CRPS are the most widely used measures for accuracy of probabilistic forecasts. Hence, their skill scores, the Brier skill score (BSS) and the mean continuous ranked probability skill score (CRPSS), are the most widely used among the skill scores in probabilistic verification. The BS and mean CRPS of a perfect forecast are zero, and hence their respective skill scores are given by Equation 2. Skill scores are particularly useful when comparing forecasts across different hydroclimatic regimes and forecast locations, as they measure the increase in accuracy of a forecast due to the "smarts" (that is, predictive skill) of the forecast system and not simply because the outcome is easier to predict (that is, the predictand has larger predictability). Care must be exercised, however, in interpreting certain skill scores without hydrological context, as explained in different parts of these guidelines.

The remainder of this chapter describes the attributes associated with the CR and LBR decompositions of the BS for probabilistic forecasts of binary events such as flooding and no flooding. As noted above, the BS is a MSE in probability space. Hence, the CR and LBR decompositions described for single-valued forecasts still apply. However, the assessment and interpretation of the attributes for probabilistic forecasts require some care and familiarization, as the forecasts and verifying observations are no longer in physical space, and the decompositions are specific to the definition of the binary events.

3.4.5 Reliability

For probabilistic forecasts, reliability answers questions such as: When the forecast says 80% probability of flooding, does flooding actually occur 80% of the time? If this is the case for all possible forecast probabilities of exceedance, the probabilistic forecasts are reliable. Formally, probabilistic forecasts are reliable, or conditionally unbiased in the type I error sense, if the forecast events are observed as frequently on average as indicated by the forecast probabilities. In such a case, the reliability component of the BS (that is, REL in the CR decomposition equation) vanishes. For example, probability forecasts for flooding of 0.2 are reliable at this specific level of forecast probability if flooding is observed 20% of the time whenever the forecast indicates a 20% chance of flooding.

If the probability forecasts are similarly reliable at all levels of forecast probability, the forecasts are said to be "well calibrated". The expression stems from the fact that, given a large enough sample size and the assumption of stationarity, probabilistic forecasts can always be made reliable via postprocessing or calibration. Calculation of reliability measures typically involves binning the forecasts into non-overlapping subranges within the dynamic range of the forecast and counting the verifying observations for each bin (see Chapter 4).

Figures 8(a) and (b) show two examples of reliable ensemble forecasts at specific levels of forecast probability. To keep the examples as simple as possible, this chapter only considers observed outcomes of flooding and no flooding. One may apply, however, different thresholds or event definitions with no loss of generality as long as the events are mutually exclusive and collectively exhaustive. Figure 8(a) and (b) are for forecasts of 80% and 20% chance of flooding, respectively. If the forecasts are reliable, one would expect 80% and 20% of the verifying observations to report flooding. Figure 8(a) and (b) show that such is indeed the case, as eight and two out of ten verifying observations exceed the flood level, respectively.

Figure 8(c) combines simpler renditions of Figure 8(a) and (b) into a single plot and is shown to illustrate how multiple reliable ensemble forecasts may collectively appear versus verifying observations in physical space. Whereas Figure 8(a) and (b) are based only on a single threshold (that is, the flood level), reliability is usually assessed with respect to multiple thresholds that encompass low to high flows for verification of ensemble streamflow forecasts. If all ensemble forecasts are reliable at all thresholds and their box-and-whisker plots are combined into a single plot, one may expect them to form a broad cluster around the one-to-one line, similar to the parallelogram in Figure 3. Figure 8(c) helps visualize how such a plot may look.

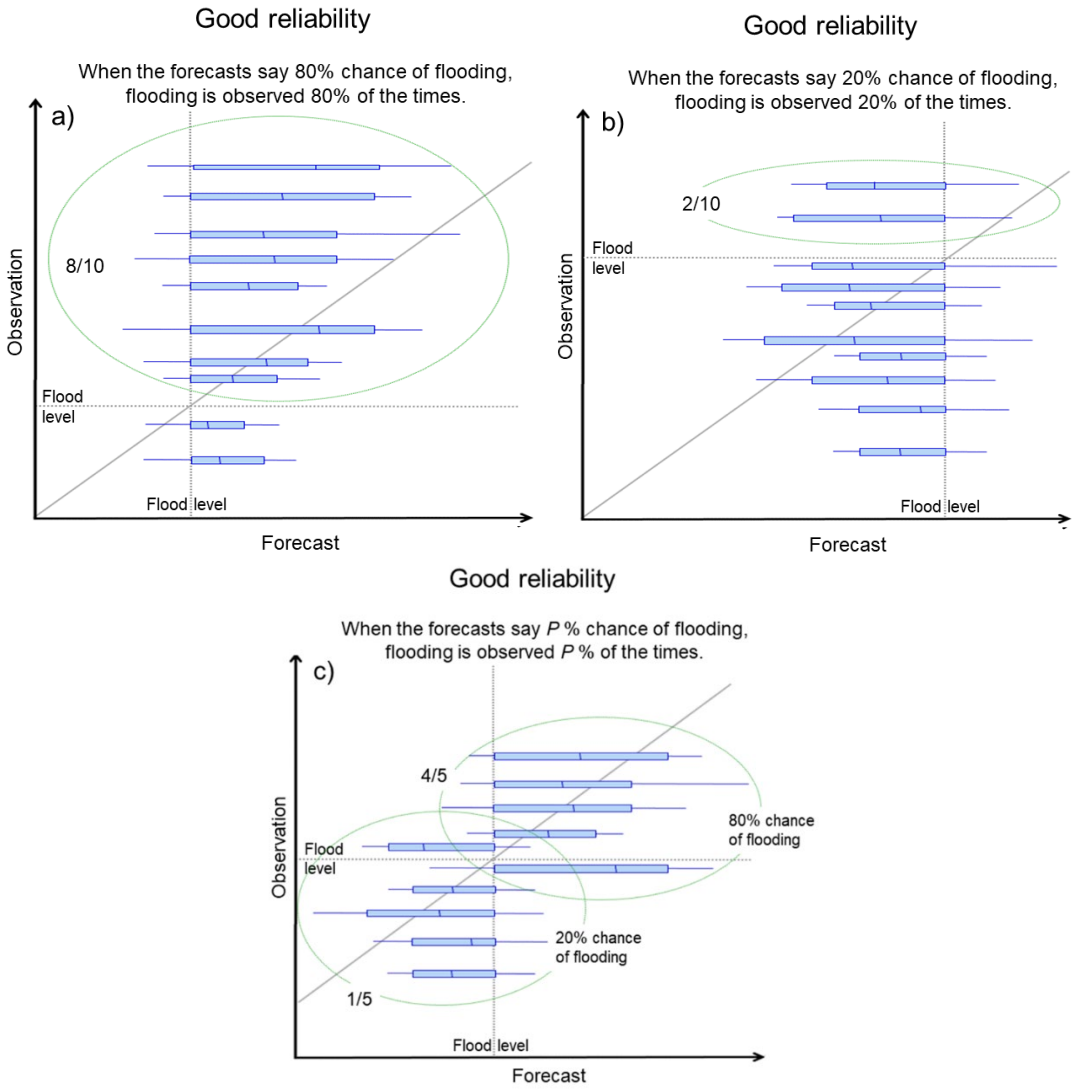


Figure 8. Illustrative examples of perfectly reliable ensemble forecasts for forecast probability of (a) 0.8 and (b) 0.2; (c) combination of simplified renditions of (a) and (b).

3.4.6 Resolution

Resolution measures the forecast's ability to differentiate different observed outcomes. Probability forecasts for flooding versus no flooding have good resolution if the forecasts of different probabilities of flooding differentiate the verifying observations of flooding from those of no flooding. Figure 9 shows examples of ensemble forecasts with good and poor resolution. In Figure 9(a), there are two forecast probabilities of flooding, 80% and 20%. For the forecast probability of flooding of 80%, four out of five verifying observations report flooding (that is, the observed frequency of flooding is 80%). For the forecast probability of flooding of 20%, one out of five verifying observations reports flooding (that is, the observed frequency of flooding is 20%). The outcome of 80% of the verifying observations reporting flooding is

significantly different from that of 20%. Hence, the probability forecasts in Figure 9(a) have good resolution. Note that, in this example, the observed frequencies of flooding are the same as the forecast probabilities. Hence, the probability forecasts in Figure 9(a) are also reliable.

In Figure 9(b), the forecast probabilities of flooding are the same as those in the first example. However, for both forecast probabilities, two out of five verifying observations report flooding (that is, the observed frequency of flooding is the same at 40% despite the large difference in the forecast probabilities). Hence, the ensemble forecasts in Figure 9(b) have poor resolution. Recall in section 3.3 that the scatters of single-valued forecasts with no resolution have a slope of zero versus verifying observations. A similar interpretation applies in Figure 9 in that the box-and-whisker plots of ensemble forecasts with no resolution tend to form a cluster with no slope.

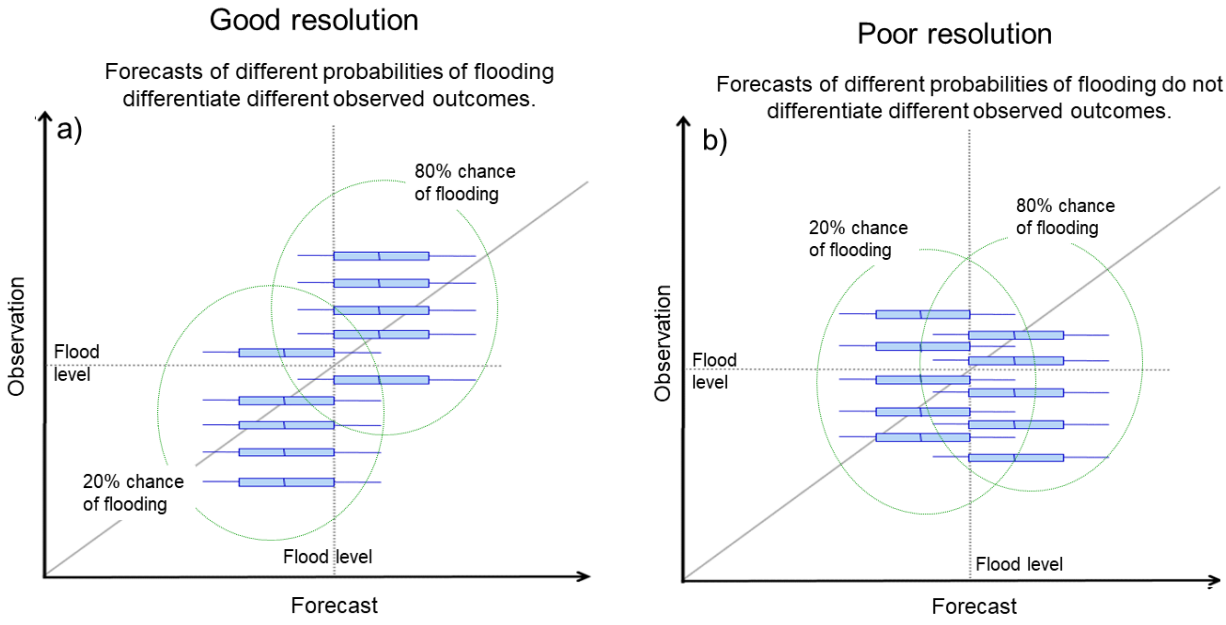


Figure 9. Illustrative examples of ensemble forecasts with (a) good and (b) poor resolution

3.4.7 Uncertainty

For probabilistic forecasting of flooding or no flooding, UNC is given by the variance of a Bernoulli random variable (for example, a coin toss), that is, $freq\,(1-freq)$, $0 \leq freq \leq 1$, where $freq$ is the observed exceedance probability or frequency (Drake, 1967). Hence, UNC is at maximum when $freq = 0.5$ and vanishes as $freq$ approaches 1 or 0. The above implies that, the higher the flood level is, the smaller the UNC component of the BS is. For probabilistic forecasts, this "smallness" of UNC for rare events has no connection to the physical world but simply reflects the fact that, with a high flood level, almost all verifying observations are for no flooding, and hence the variability of the observed outcome is extremely small in probability space. Figure 10(a) and (b) show illustrative examples of observations with large and small uncertainty in their indicator representation, respectively.

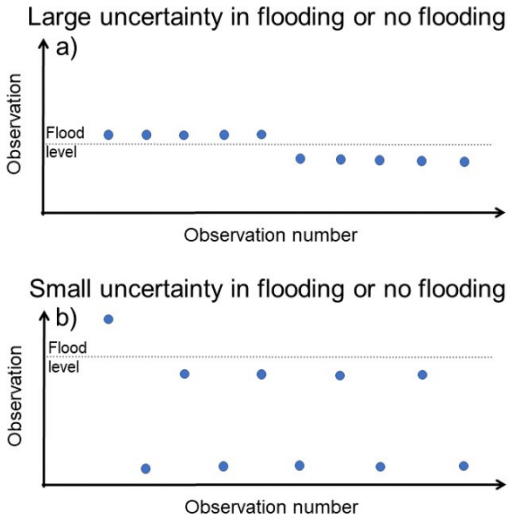


Figure 10. Illustrative examples of observations with (a) large and (b) small uncertainty in the binary outcome of flooding or no flooding

3.4.8 Type II conditional bias

For probability forecasts of flooding versus no flooding, type II conditional bias measures the departure or difference of the average forecast probabilities of flooding and no flooding from the indicator representation of the verifying observed events of flooding and no flooding (that is, 1 and 0, respectively). If the differences are small, the forecasts have small type II conditional bias, and if the differences are large, the forecasts have large type II conditional bias. Unlike reliability, type II conditional bias is not calibratable because, having failed to detect the events, the forecasts contributing to type II conditional bias are null.

Figure 11 shows examples of ensemble forecasts with small and large type II conditional bias. In Figure 11(a), the average forecast probabilities of flooding for the two groups of five ensemble forecasts associated with the observed event of flooding and no flooding are very close to 1 and 0, respectively. Hence, the ensemble forecasts in Figure 11(a) have small type II conditional bias. In Figure 11(b), the situation is reversed, and the ensemble forecasts are severely conditionally biased in the type II error sense.

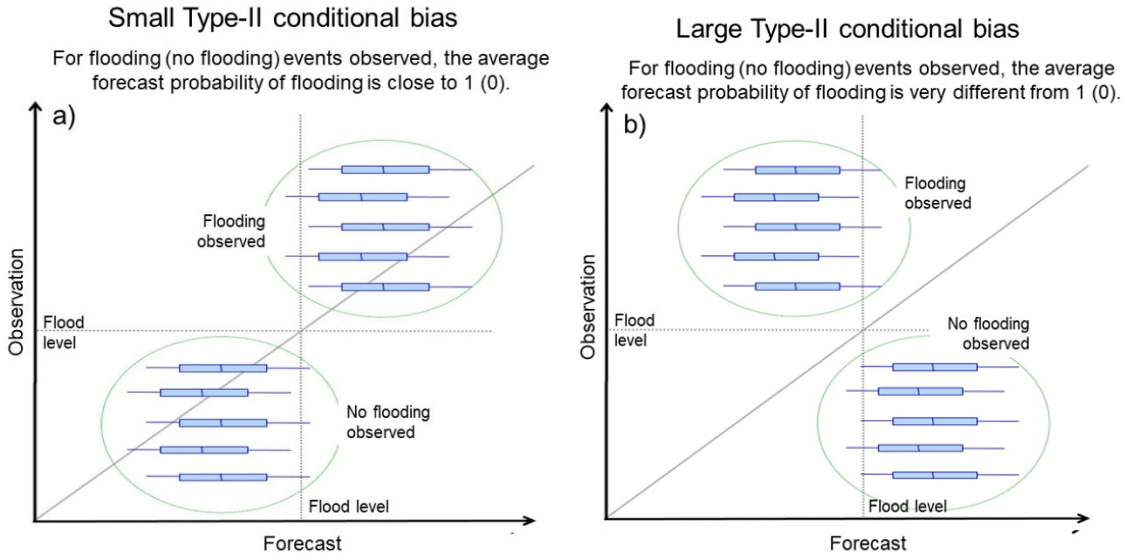


Figure 11. Illustrative examples of ensemble forecasts with (a) small and (b) very large type II conditional bias

3.4.9 Discrimination

Probabilistic forecasts for flooding versus no flooding are discriminatory if the average forecast probabilities differ significantly, with no regard to their accuracy, between the observed events of flooding and no flooding. Climatological forecasts are not discriminatory because the forecast is the same regardless of the observed event. Discrimination is measured by the differences between the (unconditional) average forecast probability of flooding and the average forecast probabilities associated with the observed events of flooding and no flooding. Measures of discrimination average the squared difference over the observed events of flooding and no flooding. Calculation of discrimination measures typically involve binning the verifying observations into the flooding and no flooding categories and sorting the forecasts into the respective bins (see Chapter 4).

Figure 12 shows examples of good and poor discrimination. In Figure 12(a), the average forecast probabilities differ significantly between the ensemble forecasts associated with the verifying observations of flooding and those of no flooding, thus exhibiting good discrimination. In Figure 12(b), the average forecast probabilities differ little between the two groups of ensemble forecasts, exhibiting no discrimination. Similarly to single-valued forecasts (see section 3.3), the fact that the box-and-whisker plots align vertically (that is, the slope is infinitely large) is an indication that the forecasts lack discrimination.

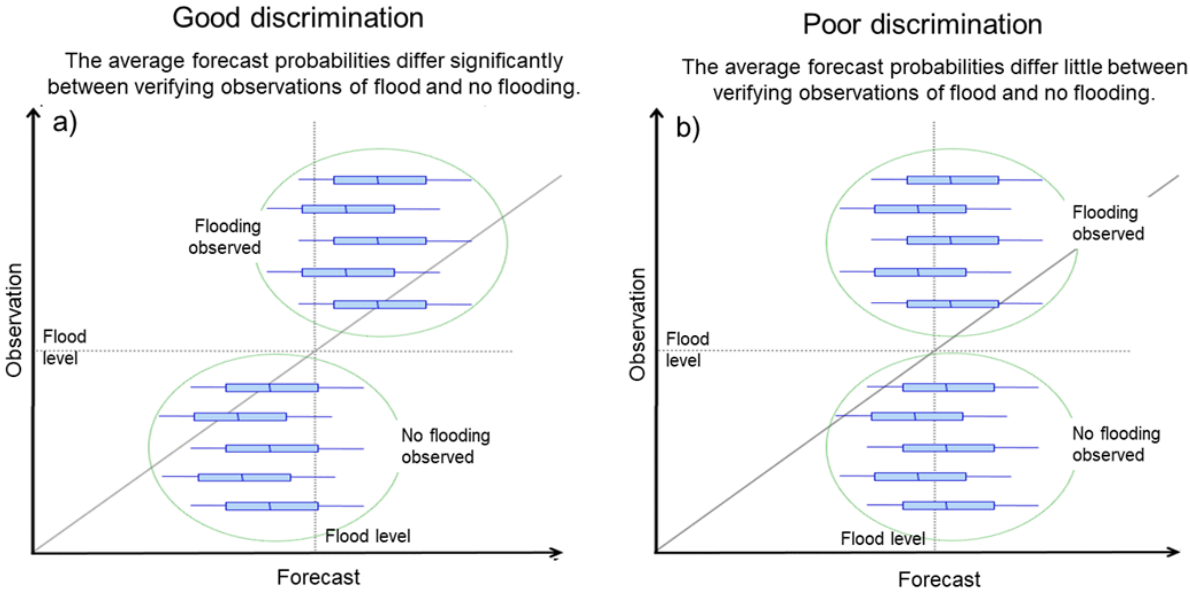


Figure 12. Illustrative examples of ensemble forecasts with (a) good and (b) poor discrimination

3.4.10 Sharpness

For probability forecasts, sharpness measures the tendency to predict with probabilities close to 0 or 1 (that is, "stick its neck out" by having all or most ensemble members predict either flooding or no flooding in the case of ensemble flood forecasting). Probability forecasts are said to be sharp if they tend to issue probabilities close to 0 or 1. A high degree of sharpness is desirable only if it improves the overall forecast accuracy. Hence, the merit of sharpness depends on the assessment of other attributes. With T2B and DIS being equal, a sharp forecast is preferred to an unsharp forecast, since sharpness improves accuracy.

Figure 13(a) and (b) shows examples of unsharp and sharp ensemble forecasts, respectively. In Figure 13(a), all ensemble forecasts indicate approximately 50% chance of flooding (or no flooding), thus exhibiting poor sharpness. In Figure 13(b), each ensemble forecast indicates a near-100% or 100% chance of either flooding or no flooding. These forecasts are extremely confident (correctly or not) about flooding versus no flooding and hence are very sharp.

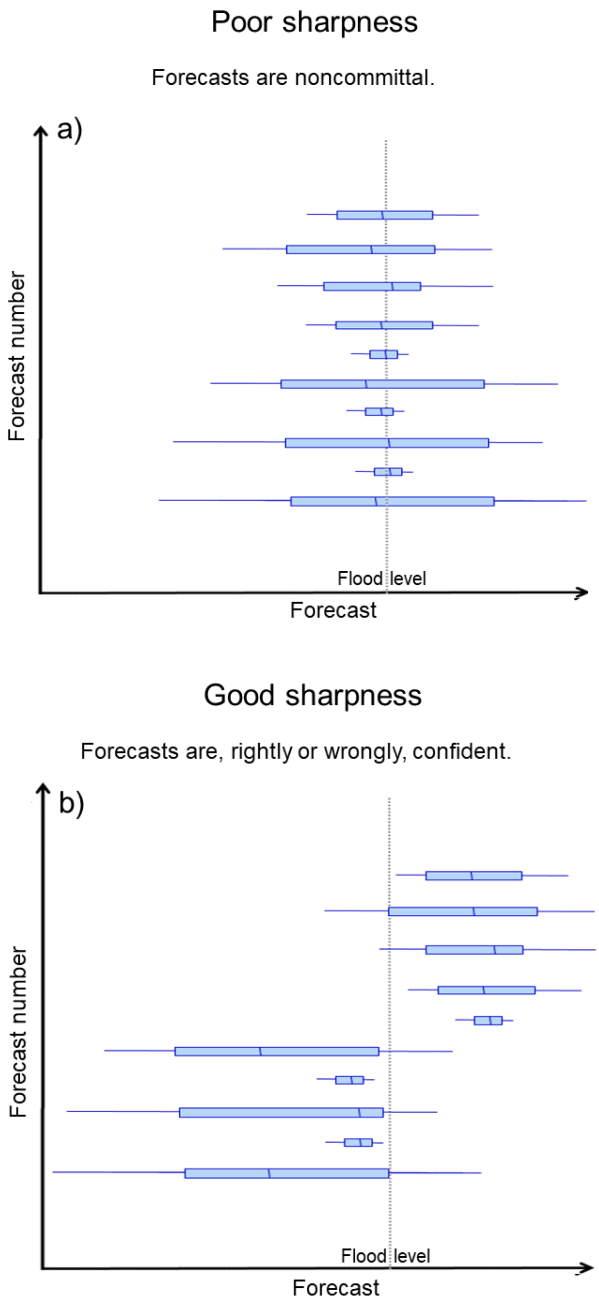


Figure 13. Illustrative examples of (a) unsharp and (b) sharp ensemble forecasts

In practice, multiple forecast attributes are generally necessary to assess forecast quality. Depending on the application and the user's risk perception and aversion, some attributes may be more important than others. Regardless of the application, quality control of the verification dataset is a prerequisite for verification. Visualization of forecasts and verifying observations using time series, scatterplots or box-and-whisker plots is often extremely helpful in identifying data issues, understanding the nature of the forecast, and recognizing visible weaknesses in the forecast (for example, in high flow conditions). The EVS provides multiple variations of the box-and-whisker plot representation of ensemble forecasts for this purpose. Certain aspects of forecast quality may be analysed visually first for qualitative assessment (similarly to the conceptual illustrations used in this chapter). Specific aspects may then be analysed numerically with verification measures for quantitative assessment. For the latter, Chapter 4 presents the commonly used verification metrics and their graphical representations.

3.5 Key points

- Forecast quality is assessed by comparing forecasts with verifying observations (or high-quality estimates) under the assumption that they are realizations of IID random variables. The relationship between the two is then described wholly by their joint distribution. Multiple attributes of forecast quality are necessary to describe the essence of this distribution.

- Accuracy describes the overall level of agreement between the forecasts and their verifying observations and hence is most representative of forecast quality. Measures of accuracy, such as the RMSE and mean CRPS for single-valued and probabilistic forecasts, respectively, reflect multiple attributes that are largely independent of one another in information content, such as correlation and biases in the mean and standard deviation in the case of the RMSE.

- Skill describes the relative accuracy of the subject forecast in comparison with a reference forecast or benchmark of choice. The reference forecast may be climatology, persistence or a forecast produced from a baseline forecast system. Skill scores calculate percent improvement in the accuracy metrics of choice by the subject forecast over the reference forecast. The Nash–Sutcliffe efficiency, which is very widely used in calibration of hydrological models, is an example of the MSE skill score.

- Several attributes of probabilistic forecasts arise from decomposing the joint probability distribution of the forecast and the verifying observation into the conditional and marginal distributions.

- In probabilistic verification, reliability (or type I conditional bias), resolution and uncertainty arise from conditioning on the forecast via the calibration–resolution (CR) decomposition, whereas type II conditional bias, discrimination and sharpness arise from conditioning on the observation via the likelihood–base rate (LBR) decomposition.

- Though verification and prediction are different, it is helpful to relate the CR decomposition with forward (that is, regular) linear regression and the LBR decomposition with reverse regression (that is, regression with the predictor and the predictand interchanged). It is also helpful to consider the CR and LBR decompositions as characterizing forecast quality from the perspective of reducing false alarms (crying wolf when there are none) and misses (failing to see the wolf), respectively, given the same absolute accuracy in the forecast.

- Reliability, resolution, type II conditional bias and discrimination are competing attributes given the absolute accuracy of the forecast. Specifically, reducing type I and type II conditional biases is a zero-sum game unless absolute accuracy is improved. Hence, assessment of individual forecast attributes is critical to assessing the trade-offs, guiding improvements in forecast systems and processes, and improving application-specific decisions based on the user's risk perception and tolerance.

- In the CR decomposition, uncertainty reflects predictability of the variable being verified. In the LBR decomposition, sharpness measures the forecast's ability to "stick its neck out", correctly or incorrectly. Though these two attributes do not pertain to the joint relationship between forecast and observation, they contribute to the overall accuracy and hence should be assessed. When assessing uncertainty, it is a good practice to consider skewness (asymmetry in distribution) and heteroscedasticity (nonuniformity in variability) to aid possible stratification, pooling or conditioning of the forecast–observation pairs.

- The above points regarding probabilistic verification mean that, between reliability and resolution and between type II conditional bias and discrimination, it is generally necessary to assess only one of the two attributes in each pair. Commonly, the choices are reliability and discrimination.

CHAPTER 4. COMMONLY USED VERIFICATION METRICS

A wide range of verification metrics have emerged in the atmospheric sciences (Jolliffe and Stephenson, 2012; Wilks, 2011; Casati et al., 2008) and in other disciplines (Stephenson and Jolliffe, 2003). More recently, existing verification metrics have been adapted or newly developed to meet specific needs in hydrology and water resources applications and to provide meaningful verification results to specific user groups (see, for example, Brown et al., 2010; Liu et al., 2011; Zappa et al., 2013). Anctil and Ramos (2019) present a wide range of examples of hydrological verification, highlighting the various objectives, forecast–observation datasets and verification metrics reported in the literature. The Joint Working Group on Forecast Verification Research of the World Weather Research Programme (WWRP) and the Working Group on Numerical Experimentation maintains a reference website describing the standard and newly-developed verification metrics: https://www.cawcr.gov.au/projects/verification/.

Many of the metrics may be found in verification software such as the Ensemble Verification System (EVS) (Brown et al., 2010) (see https://sourceforge.net/projects/ensemble-verification-system/). Developed originally by the United States National Weather Service (NWS) in support of operational hydrological forecasting, the EVS is a freely available open-source software tool for verification of ensemble forecasts of hydrological and hydrometeorological variables such as streamflow, precipitation and temperature. One may also use the EVS for verification of single-valued forecasts as one-member ensemble forecasts (see Cases 1 and 2 in Chapter 7). The EVS includes detailed documentation of a comprehensive set of verification metrics, including new and more application-oriented additions. As indicated in Tables 1 and 2, Chapter 7 and Appendix A present several case studies with interpretations of the verification metrics and diagrams. The case studies include hands-on examples using the EVS, Python packages, R and MATLAB scripts.

4.1 Introduction

An array of widely used verification metrics is applied in hydrological verification of single-valued (deterministic) and probabilistic forecasts. The former may be deterministic forecasts or single-valued reductions of ensemble forecasts. The latter may be in the form of ensembles or probability distributions. Table 5 lists the verification metrics commonly used in operational water and weather forecasting as grouped by forecast quality attribute. Given the verification task, one may determine the forecast type, identify the forecast attributes to be assessed and select the verification metrics from the table.

Table 5. Verification metrics commonly used in operational water and weather forecasting as grouped by forecast quality attribute

Attribute	*Metric name*	*Type of forecast*	*Discrete events?*
Accuracy	Mean absolute error (MAE)	Single-valued	No
	Mean squared error (MSE)	Single-valued	No
	Root mean squared error (RMSE)	Single-valued	No
	Critical success index (or threat score)	Single-valued	Yes
	Mean continuous ranked probability score (CRPS)	Probabilistic	No
	Brier score (BS)	Probabilistic	Yes
	Ranked probability score (RPS)	Probabilistic	Yes

Attribute	*Metric name*	*Type of forecast*	*Discrete events?*
Bias (first order)	Relative mean error (or relative bias)	Single-valued	No
	Frequency bias (FB)	Single-valued	Yes
	Mean error (ME)	Single-valued	No
Correlation	Pearson correlation coefficient (CORR)	Single-valued	No
	Spearman rank correlation	Single-valued	No
Skill	Mean absolute error skill score (MAESS)	Single-valued	No
	Mean squared error skill score (MSESS)	Single-valued	No
	Equitable threat score (or Gilbert skill score)	Single-valued	Yes
	Mean continuous ranked probability skill score (CRPSS)	Probabilistic	No
	Brier skill score (BSS)	Probabilistic	Yes
	Ranked probability skill score (RPSS)	Probabilistic	Yes
Reliability (type I conditional bias)	Mean squared error reliability	Single-valued	No
	Success ratio	Single-valued	Yes
	Mean CRPS reliability	Probabilistic	No
	Brier score reliability	Probabilistic	Yes
	Reliability diagram	Probabilistic	Yes
	Rank histogram (or Talagrand diagram)	Probabilistic	Yes
	Spread-bias diagram	Probabilistic	No
Resolution	Mean squared error resolution	Single-valued	No
	Mean CRPS resolution	Probabilistic	No
	Brier score resolution	Probabilistic	Yes
Type II conditional bias	Mean squared error type II conditional bias	Single-valued	No
	Brier score type II conditional bias	Probabilistic	Yes
Discrimination	Mean squared error discrimination	Single-valued	No
	Probability of detection (or hit rate)	Single-valued	Yes
	Probability of false detection (or false alarm rate)	Single-valued	Yes
	Brier score discrimination	Probabilistic	Yes
	Relative operating characteristic (ROC) curve	Both	Yes
	ROC score	Both	Yes
Sharpness	Forecast frequency histogram	Probabilistic	Yes
	Average width of the prediction intervals	Probabilistic	No
	Variance	Both	No
	Standard deviation	Both	No
Uncertainty	Variance	Both	Yes
	Standard deviation	Both	Yes

The rest of this chapter is organized as follows. Sections 4.2 and 4.3 present the commonly used metrics for verification of discrete and continuous variables, respectively. The above sections also describe the metrics pertaining to the CR and LBR decompositions of the MSE and BS, and the CR decomposition of the mean CRPS. Section 4.4 discusses several practical considerations regarding selection of suitable metrics, intercomparison of verification results for different forecast locations and under different conditions, assessing sampling uncertainty associated with the verification statistics, and conditional verification for the assessment of flow regime-specific forecast quality.

4.2 Metrics for categorical forecasts

The verification metrics for forecasts of discrete events are presented in this section. This type of forecast has a limited number of possible outcomes. For streamflow forecasting, the simplest examples include forecasts for flooding and no flooding (applying a single threshold to streamflow), and those for major, minor and no flooding (applying two thresholds). When only a single threshold is used, the observed outcome is binary: either flooding is observed or not. When the observation is represented as an indicator variable, the outcome is 1 and 0 for flooding and no flooding, respectively. For single-valued forecasts, the categorical forecast is also binary: either flooding is forecast if the forecast exceeds the threshold, or no flooding is forecast if it does not exceed the threshold. For probabilistic forecasts, the forecast specifies the probability of exceeding the threshold, p, and the probability of not exceeding the threshold, $1 - p$.

4.2.1 Scores derived from the contingency table

Contingency tables are used to describe the discrete joint distribution of forecast and observation in terms of the frequencies of occurrence for defined categories. For verification of binary forecasts, the 2 × 2 contingency table is defined as shown in Table 6.

Table 6. Definition of the 2 × 2 contingency table for the verification of binary events (for example, flooding or no flooding)

2 × 2 contingency table		Event observed		Total
		Yes	No	
Event forecast	Yes	H (hits)	FA (false alarms)	H + FA
	No	M (misses)	TN (true negatives)	M + TN
Total		H + M	FA + TN	H + M + FA + TN

In Table 6, the rows and columns represent the forecast and observed categories, respectively. The yes- and no-events represent flooding and no flooding, respectively, for both observations and forecasts. For probabilistic forecasts, the table is derived for a given probability level to categorize each probability forecast as a yes-event if the forecast probability exceeds the chosen probability level, or a no-event otherwise. The 2 × 2 contingency table includes the following entries:

- **Hits** (H) indicates the number of true positives, that is, observed yes-events that were correctly forecast as yes-events;
- **False alarms** (FA) indicates the number of false positives, that is, observed no-events that were incorrectly forecast as yes-events;

- **Misses** (M) indicates the number of false negatives, that is, observed yes-events that were incorrectly forecast as no-events;

- **True negatives** (TN) indicates the number of observed no-events that were correctly forecast as no-events.

The total number of observed yes- and no-events are H + M and FA + TN, respectively. The total number of forecast yes- and no-events are H + FA and M + TN, respectively. A number of verification scores may be derived from the contingency table as described below.

The **probability of detection** (POD), also known as the hit rate, is a measure of discrimination representing the proportion of the correctly forecast yes-events (hits) among all observed yes-events. For the 2 × 2 contingency table, the POD is given by:

$$POD = \frac{H}{H + M} \tag{6}$$

The POD ranges from 0 (all observed events are misses) to 1 (all observed events are hits); a perfect score is 1.

The **probability of false detection** (POFD), also known as the false alarm rate, is a measure of discrimination representing the proportion of the incorrectly forecast no-events (false alarms) among all observed no-events. The POFD is defined as the number of false alarms divided by the total number of observed no-events. For the 2 × 2 contingency table, the POFD is given by:

$$POFD = \frac{FA}{FA + TN} \tag{7}$$

The POFD ranges from 0 (no false alarms) to 1 (all observed no-events are false alarms); a perfect score is 0.

The **success ratio** (SR) is a measure of reliability representing the proportion of the correctly forecast yes-events (hits) among all forecast yes-events. The SR is defined as the number of hits divided by the total number of forecast yes-events. For the 2 × 2 contingency table, the SR is given by:

$$SR = \frac{H}{H + FA} \tag{8}$$

The SR ranges from 0 (all forecast yes-events are false alarms) to 1 (no false alarms); a perfect score is 1. The categorical performance diagram (Roebber, 2009) plots the POD as a function of the SR.

The **frequency bias** (FB) is a measure of bias defined as the ratio of the total number of forecast yes-events divided by the total number of observed yes-events. For the 2 × 2 contingency table, the FB is given by:

$$FB = \frac{H + FA}{H + M} = \frac{POD}{SR} \tag{9}$$

The FB ranges from 0 to infinity; a perfect score is 1. Values lower and higher than 1 indicate tendency for under-forecasting (too many misses) and over-forecasting (too many false alarms), respectively.

The **false alarm ratio** (FAR) is a measure of reliability representing the proportion of the forecast yes-events that are false alarms. The FAR is defined as the number of false alarms divided by the total number of forecast yes-events. For the 2 × 2 contingency table, the FAR is given by:

$$FAR = \frac{FA}{H + FA} = 1 - SR \tag{10}$$

The FAR ranges from 0 (no false alarms) to 1 (all forecast yes-events are false alarms); a perfect score is 0.

The **fraction correct** (FC) represents the proportion of correct forecasts and is defined as the number of hits and true negatives divided by the total number of events. For the 2 × 2 contingency table, the FC is given by:

$$FC = \frac{H + TN}{H + M + FA + TN} \tag{11}$$

The FC ranges from 0 to 1; a perfect score is 1. The FC is influenced by the number of true negatives. For rare yes-events with a very large number of no-events that are trivially easy to forecast most of the time, the score will be close to 1 owing to a very large number of true-negatives (see the Finley affair in Chapter 2). For example, in the dry season in an arid region, there is only a miniscule chance of extreme flooding on any given day. Hence, by forecasting no extreme flooding every day, one can score very highly on the FC, even though such a technique has no chance of ever scoring a single hit. A better alternative, especially for rare events, is the CSI score, which accounts for hits, misses and false negatives.

The **critical success index** (CSI), also called the threat score, represents the proportion of the correctly forecast yes-events among all yes-events, forecast or observed, and is defined as the number of hits divided by the number of hits, misses and false alarms combined. For the 2 × 2 contingency table, the CSI is given by:

$$CSI = \frac{H}{(H + M + FA)} = \frac{1}{\frac{1}{SR} + \frac{1}{POD} - 1} \tag{12}$$

The CSI ranges from 0 (no hits) to 1 (no misses and no false alarms); a perfect score is 1. The CSI may be interpreted as a measure of forecast accuracy without considering the true negatives.

As some hits could occur due to random chance, a modified score known as the equitable threat score has also been defined.

The **equitable threat score** (ETS), also called Gilbert skill score, represents the proportion of the correctly forecast yes-events over all yes-events, forecast or observed, adjusted for the hits due to random chance (one is more likely to correctly forecast the more frequently occurring events). The ETS is defined as the number of hits minus the number of hits due to random chance, divided by the sum of the number of hits, misses, and false alarms combined minus the number of hits due to random chance. For the 2 × 2 contingency table, the ETS is given by:

$$ETS = \frac{(H - H_{random})}{(H + M + FA - H_{random})} \tag{13}$$

where

$$H_{random} = \frac{(H + M) \times (H + FA)}{(H + M + FA + TN)} \tag{14}$$

The ETS ranges from –1/3 to 1, a value below 0 indicating no skill; a perfect score is 1. By considering hits due to random chance, the ETS allows fairer assessment of the skill of forecasts across different regimes.

The **Peirce's skill score** (PSS, also called Hanssen and Kuipers discriminant) is a measure of skill in forecasting the yes- and no-events considering all four bins in the contingency table and is defined as the difference between the POD and the POFD. For the 2 × 2 contingency table, the PSS is given by:

$$PSS = POD - POFD \tag{15}$$

The PSS ranges from –1 to 1, a value below 0 indicating no skill; a perfect score is 1 (no misses and no false alarms).

The **base rate (BR)** describes the rate of occurrence of the observed yes-events, and is defined as:

$$Base\ rate = \frac{number\ of\ observed\ yes}{total\ number\ of\ events} = \frac{(H + M)}{(H + M + FA + TN)} \tag{16}$$

The BR is purely a characteristic of the observations, not the forecasts being evaluated, and varies between 0 and 1.

The **probability of forecast of occurrence** (POFO) is equivalent to the BR but for the forecast yes-events, and is defined as:

$$POFO = \frac{number\ of\ forecast\ yes}{total\ number\ of\ events} = \frac{(H + FA)}{(H + M + FA + TN)} \tag{17}$$

The POFO ranges between 0 and 1.

Contingency tables can also be defined with more than two categories. For example, the 3 × 3 contingency table based on two different flood thresholds (*Threshold1* and *Threshold2*) includes the following events for both the forecast and the observation:

- {streamflow < *Threshold1*}
- {*Threshold1* ≤ streamflow < *Threshold2*}
- {streamflow ≥ *Threshold2*}

Each of the metrics derived from the contingency table may be defined for a specific category. Wilks (2011) details converting a 3 × 3 contingency table into three different 2 × 2 contingency tables for each of the three categories and deriving the different contingency scores for each category.

4.2.2 Scores for probabilistic categorical forecast

There are two commonly used scores for verification of probabilistic categorical forecasts: the Brier score (BS) for binary events and the average ranked probability score (RPS) for categorical events. The original formulation of the BS (Brier, 1950) is applicable to categorical forecasts as well. In this document, however, only the binary form of the BS, sometimes referred to as the half BS, is used throughout.

The **Brier score** (BS) is a MSE of probability forecasts for binary events and thus measures the accuracy of probability forecasts for binary events:

$$BS = \frac{1}{n}\Sigma_{k=1}^{n}(F_k - O_k)^2 \tag{18}$$

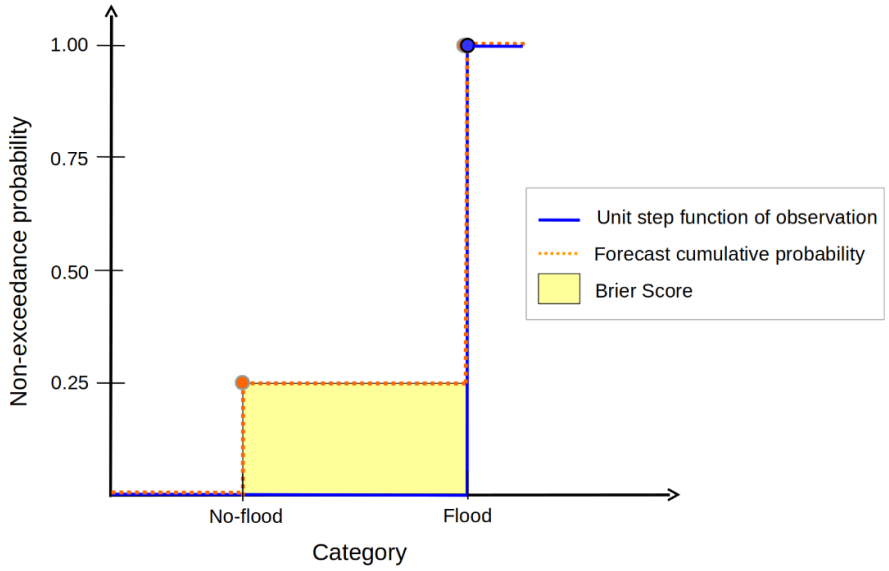
where n is the total number of forecast–observation pairs, F_k is the kth forecast probability of the occurrence of the event and O_k is the indicator representation of the verifying observed

outcome (that is, 1 if the event occurred and 0 otherwise). The BS ranges from 0 (perfect score) to 1. For example, if the probability forecast says 80% chance of discharge exceeding the flood level and flooding does occur, the squared error of the probability forecast is $(0.8 - 1.0)^2 = (-0.2)^2 = 0.04$. If flooding does not occur, the squared error of the probability forecast is $(0.8 - 0)^2 = 0.64$.

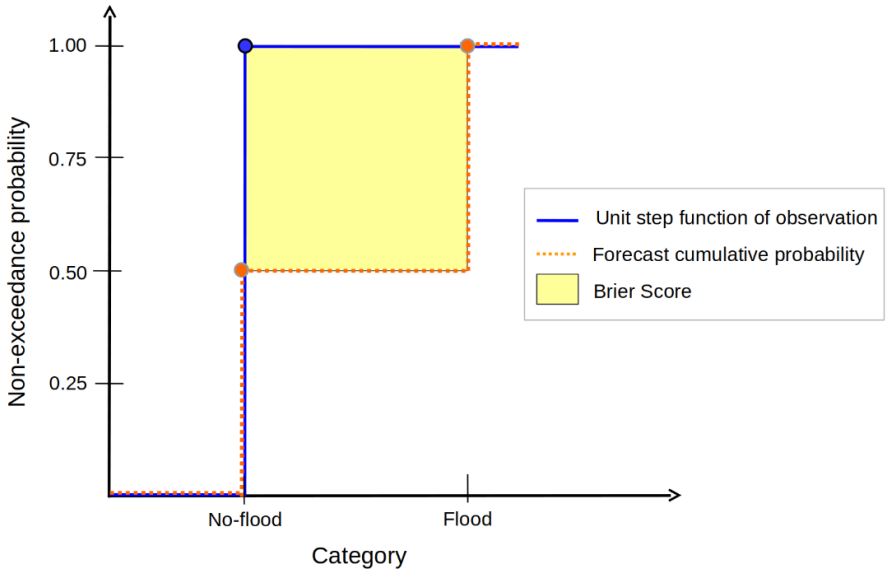
A BS of 0 indicates a perfectly accurate (and hence perfectly sharp) probability forecast (that is, when the forecast indicates a probability of flooding of 1, flooding does occur). A BS of 1 indicates a perfectly inaccurate (and perfectly sharp) forecast (that is, when the forecast indicates a probability of flooding of 0, flooding does occur). The BS is very useful when the consequences of being below and above a specific threshold, such as a flood level, are asymmetric. Figure 14 and Table 7 illustrate how to visualize and compute the BS with three different streamflow forecast–observation pairs. The two categories are defined with a flood threshold, denoted as Thr_{flood} below:

- No flooding if flow $< Thr_{flood}$
- Flooding if flow $\geq Thr_{flood}$

Forecast 1



Forecast 2



Forecast 3

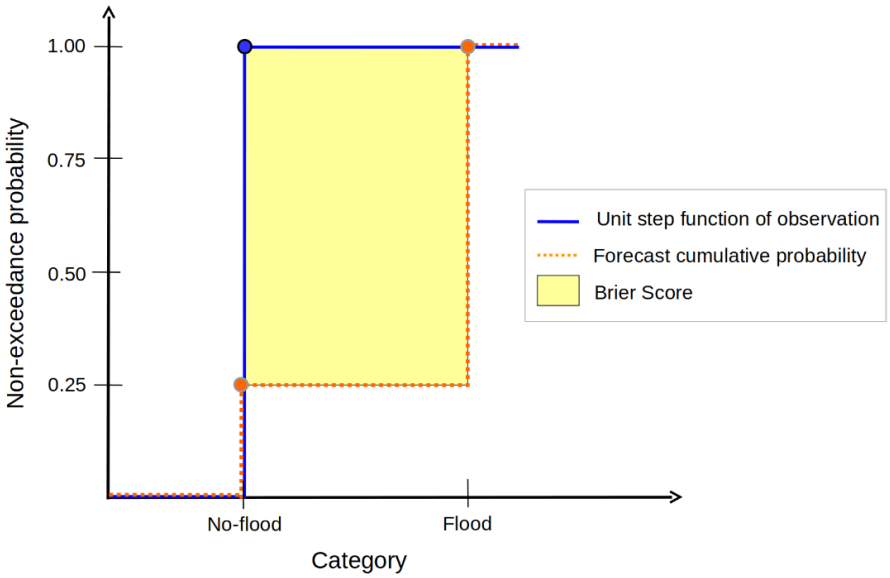


Figure 14. Illustrative examples of the BS for each of three different probability forecasts

Table 7. Computation of the BS for each of the three probability forecasts shown in Figure 14

Forecast k	Observed flow ≥ Thr_{flood}? O_k = 1 if yes, 0 otherwise	Forecast flood probability F_k = Prob [forecast flow ≥ Thr_{flood}]	BS_k of kth forecast $BS_k = (F_k - O_k)^2$
1	1	0.75	0.06
2	0	0.50	0.25
3	0	0.75	0.56
			Σ = 0.87

The BS for all three forecasts is the average of BS_k, k = 1, 2, 3;
BS = (0.06 + 0.25 + 0.56) / 3 = 0.87 / 3 = 0.29.

The components of the BS under the CR decomposition (see Equation 62 in Appendix B) may be estimated via (Murphy, 1973):

$$BS = \frac{1}{n}\Sigma_{l=1}^{L} n_l\left(\overline{O_l} - F_l\right)^2 - \frac{1}{n}\Sigma_{l=1}^{L} n_l\left(\overline{O_l} - \overline{O}\right)^2 + \overline{O}\left(1 - \overline{O}\right) \quad (19)$$

where the first, second and third terms represent REL, –RES and UNC, respectively, n is the total number of forecast probability–binary observation pairs, L is the number of bins for the forecast events that the individual forecast probabilities are subgrouped into, n_l is the number of pairs in the l th bin, $\overline{O_l}$ is the mean of the binary observations associated with the forecast probabilities in the l th bin, F_l is the l th forecast event and $\overline{O}$ is the grand mean of all binary observations. The BS is a MSE of probability forecast. Hence, the expressions above for the BS are identical to those for the MSE (see section 4.3). The only difference in the above expression is that the observations are indicator variables (that is, outcomes of a Bernoulli random variable); the observation variance (that is, the UNC term) in the BS simplifies to $\overline{O}(1 - \overline{O})$.

The components of the BS under the LBR decomposition (see Equation 63 in Appendix B) may be estimated via:

$$BS = \frac{1}{n}\Sigma_{i=0}^{1} n_i\left(\overline{F_i} - O_i\right)^2 - \frac{1}{n}\Sigma_{i=0}^{1} n_i\left(\overline{F_i} - \overline{F}\right)^2 + \frac{1}{n}\Sigma_{k=1}^{n}\left(F_k - \overline{F}\right)^2 \quad (20)$$

where the first, second and third terms represent T2B, –DIS and SHA, respectively, n is the same as in Equation 19, n_i is the number of pairs in the i th bin of observations of flooding ($i = 0$) and no flooding ($i = 1$), $\overline{F_i}$ is the mean of the forecast probabilities associated with the observations in the i th bin, O_i is the i th observed event, $\overline{F}$ is the grand mean of all forecast probabilities and F_k is the k th forecast probability. The above expression is identical to that of the MSE (see section 0). The only notational difference is that, for the BS, the index i ranges only from 0 to 1 for the observed events of 0 and 1, respectively.

In the above, the SHA term is given by the variance of the forecast, whereas in section 3.4, sharpness for probabilistic forecast is described in effect as (Daan, 1984):

$$S = \frac{1}{n}\Sigma_{k=1}^{n} F_k(1 - F_k) \quad (21)$$

In the above, S represents the average variance of n Bernoulli random variables with the probabilities of occurrence of the event prescribed by F_k, $k = 1,..,n$. If all forecast probabilities

are exclusively 0s and 1s, S is zero (or perfectly sharp), and if all forecast probabilities are 0.5 (that is, perfectly noncommittal), S is at the maximum of 0.25 (or perfectly unsharp). If the probability forecasts are unconditionally unbiased, SHA relates linearly with S (Potts, 2011. Hence, one may relate small and large S (that is, sharp and unsharp probabilistic forecast) with large and small SHA in the LBR decomposition of the BS, respectively.

The **ranked probability score (RPS)** is a multicategory extension of the binary form of the BS. The RPS measures the accuracy of categorical probability forecasts and is given by the MSE of categorical probability forecasts. For a two-category forecast, the RPS is the same as the BS. The RPS for the *k*th forecast–observed pair is given by:

$$RPS_k = \sum_{m=1}^{J} \left(F_{k,m} - O_{k,m}\right)^2 \tag{22}$$

where J is the total number of categories, $F_{k,m}$ is the forecast cumulative probability of the mth category and $O_{k,m}$ is the unit step (that is, Heaviside) function representation of the verifying observation: $O_{k,m} = 1$ for m greater than or equal to the category in which the observation falls and $O_{k,m} = 0$ otherwise. Note that $F_{k,J}$ and $O_{k,J}$ are always equal to 1 since $F_{k,m}$ is a cumulative probability and $O_{k,m}$ is bounded by 1. Therefore, their square difference $(F_{k,J} - O_{k,J})^2$ is equal to 0. The RPS ranges from 0 (perfect) to $J - 1$ (when all forecasts are incorrect). The average or mean RPS is defined as:

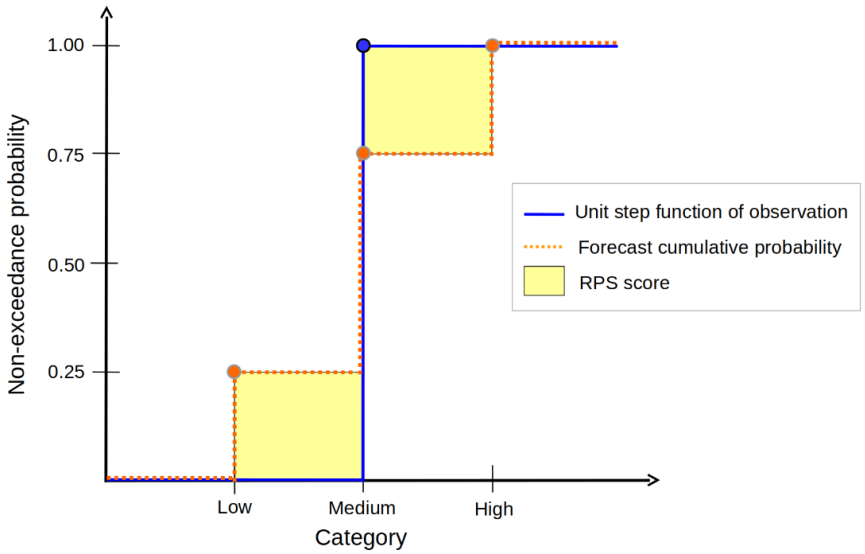
$$\overline{RPS} = \frac{1}{n}\Sigma_{k=1}^{n} RPS_k \tag{23}$$

where n is the number of forecast–observed pairs. The RPS for a given forecast–observation pair may be visualized by plotting the cumulative distribution function (CDF) and the unit step function representation of the forecast and the observation, respectively, as shown in Figure 15. The area between the CDF and the unit step function represents the RPS.

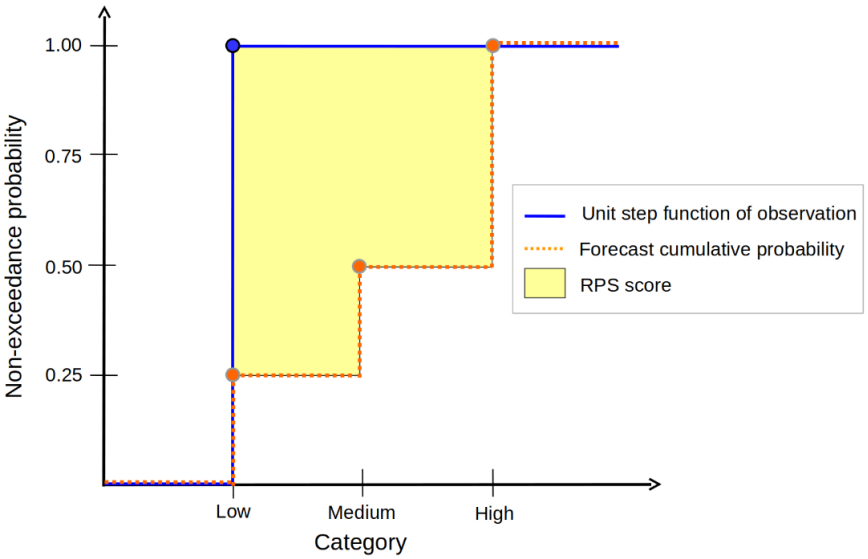
The RPS is useful for verifying categorical forecasts when one is interested more in the overall forecast accuracy across all categories than forecast accuracy for a specific category. Figure 15 and Table 8 illustrate how to visualize and compute the RPS for three different streamflow forecast–observation pairs for probability forecast with three categories. The categories are defined as:

- Low category if flow < Thr_{low}
- Medium category if Thr_{low} ≤ flow < Thr_{high}
- High category if flow ≥ Thr_{high}

Forecast 1



Forecast 2



Forecast 3

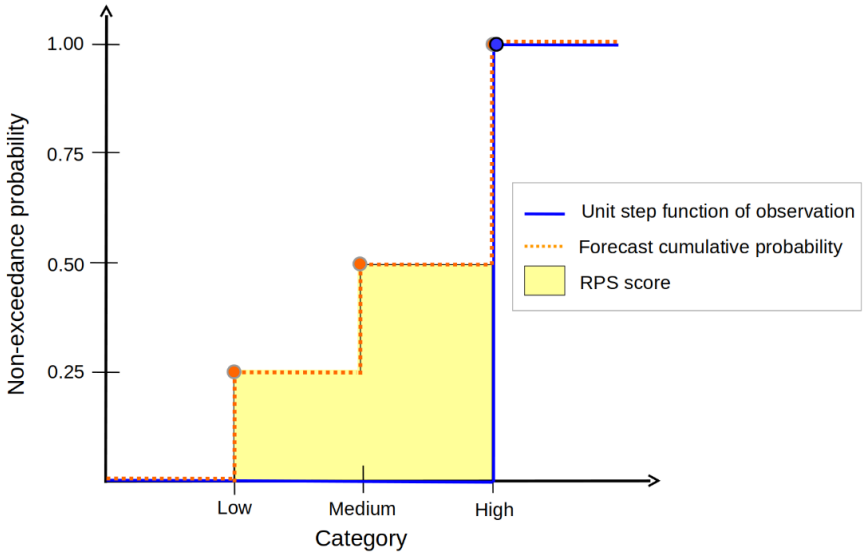


Figure 15. Illustrative examples of the RPS for three different forecast–observation pairs for probability forecast with three categories

Table 8. Computation of the RPS for the three forecast–observation pairs in Figure 15

Forecast k	Observed flow in low category? $O_{k,1}$ = 1 if yes, 0 otherwise	Observed flow in medium or lower category? $O_{k,2}$ = 1 if yes, 0 otherwise	Observed flow in high or lower category? $O_{k,3}$ = 1 if yes, 0 otherwise	Forecast probability for low category, $F_{k,1}$	Forecast probability for medium or lower category, $F_{k,2}$	Forecast probability for high or lower category, $F_{k,3}$	$RPS_k \sum_{m=1}^{3} \square \left(F_{k,m} - O_{k,m}\right)$
1	0	1	1	0.25	0.75	1	$(0 - 0.25)^2 + (1 - 0.75)^2 + (1 - 1)^2 = 0.125$
2	1	1	1	0.25	0.50	1	0.812
3	0	0	1	0.25	0.50	1	0.312
							Σ = 1.249

From Table 8, the average RPS is given by (0.125 + 0.812 + 0.312) / 3 = 1.249 / 3 = 0.42. Per Equation 2, one may define the Brier skill score (BSS) and the ranked probability skill score (RPSS) as shown below.

The **Brier skill score** (BSS) measures the relative reduction of the BS by the subject forecast over the reference forecast:

$$BSS = 1 - \frac{BS}{BS_{ref}} \tag{24}$$

where BS_{ref} is the BS of the reference forecast. The BSS ranges from −∞ to 1 (perfect score). A negative BSS value indicates that the subject forecast is worse than the reference forecast in terms of the BS, and a positive value indicates that it is better. For the example given in Table 5, let us assume that the BS for a climatological forecast is 0.33. Then, the BSS in reference to climatology is given by:

$$BSS = 1 - \frac{BS}{BS_{ref}} = 1 - \frac{0.29}{0.33} = 0.12 \tag{25}$$

The above indicates that the subject forecast is on average 12% better, in terms of the BS, than climatology. In practice, the calculation of the BS and BSS should be based on much larger forecast–observation datasets for both the subject and the reference forecasts.

The BS is a strictly proper scoring rule and hence yields the lowest (that is, the best) score when the forecaster reports the true probability (Bernoulli) distribution. Such scoring rules hence encourage the forecaster to make careful assessments and to be honest (Gneiting and Raftery, 2007). However, when using the BSS for very rare events (and hence for very frequent events as well, because a probability of 0.001 of the event occurring is the same as a probability of 0.999 of the event not occurring), additional care should be exercised (Benedetti, 2010). The issue is particularly relevant for hydrological extremes as elaborated below in the context of flood frequency analysis for illustrative purposes only.

Let us consider categorical probabilistic forecasting. The expected (half) BS for the forecast probability of p that an *M*-yr flood may occur in any given year is given by:

$$\overline{BS} = (p-1)^2 freq + (p-0)^2(1-freq) = (p-freq)^2 + freq(1-freq) \quad (26)$$

where *freq is* the observed frequency of *M*-yr flood (that is, *freq=1/M*, and the overbar signifies the expected value. Assume that the probability forecast for the above event from the old forecast system is 0, (that is, it never forecasts a flood with a return period of *M* years or larger). With much improvement in the forecast system, the new system is now perfectly reliable without even resorting to calibration (that is, *p* = *freq* for all *M* ≥ 2). The BS skill score of the new probability forecast in reference to the old is then given by:

$$BSS = 1 - \frac{\overline{BS_{new}}}{\overline{BS_{old}}} = 1 - \frac{freq(1-freq)}{freq} = freq = \frac{1}{M} \quad (27)$$

According to the above skill scoring, one might conclude, against hydrological sense, that the new forecast has improved skill over the old by 50% for 2-yr floods, by 10% for 10-yr floods, by 1% for 100-yr floods and by 0.1% for 1 000-yr floods. Note that one can predict 2-yr floods with perfect reliability by tossing a fair coin every year, whereas it will take unprecedented efforts to predict 1 000-yr floods with reliability. The above counterintuitive result arises because the BS does not penalize forecasts of zero probability (an extremely strong statement) heavily enough when they are wrong (Jewson, 2004).

The **ranked probability skill score** (RPSS) measures the relative reduction in the average RPS by the subject forecast over the reference forecast:

$$RPSS = 1 - \frac{\overline{RPS}}{\overline{RPS}^{ref}} \quad (28)$$

where $\overline{RPS}^{ref}$ is the average RPS of the reference forecast. The RPS range from −∞ to 1 (perfect score). A negative RPSS value indicates that the subject forecast is worse than the reference forecast in terms of the average RPS. For the example given in Table 6, let us assume that the average RPS for climatological forecasts is equal to 1. The RPSS in reference to climatology is then given by:

$$RPSS = 1 - \frac{\overline{RPS}}{\overline{RPS}^{ref}} = 1 - \frac{0.42}{1.0} = 0.58 \quad (29)$$

The above indicates that the subject forecast is 58% better than climatology in terms of the average RPS. As with the BSS example above, the RPSS calculation in practice should be based on much larger forecast–observation datasets for both the subject and the reference forecasts.

4.3 Metrics for continuous forecasts

This section describes verification metrics for forecasts of continuous variables, such as streamflow.

4.3.1 Scores for continuous single-valued forecasts

A number of metrics are usually used to assess the quality of single-valued forecasts.

The **mean absolute error (MAE**) is the average of the absolute differences between forecasts and observations:

$$MAE = \frac{1}{n}\Sigma_{k=1}^{n}|f_k - o_k| \tag{30}$$

where n is the number of forecast–observation pairs, and f_k and o_k are the kth forecast and observation, respectively. The perfect score for the MAE is 0. Because all errors are equally weighted, the MAE is not as sensitive to large forecast errors as the MSE (see below). The mean continuous ranked probability score (CRPS) (see below) of single-valued forecasts is equivalent to the MAE.

The **mean squared error (MSE)** is the mean squared difference between forecast and observation:

$$MSE = \frac{1}{n}\Sigma_{k=1}^{n}(f_k - o_k)^2 \tag{31}$$

The perfect score for the MSE is 0. The MSE is a measure of the overall accuracy of the forecast and comprises bias in mean, bias in standard deviation and the degree of association (that is, correlation) (see Equation 61 in Appendix B). The MSE is sensitive to large forecast errors. In hydrological applications, forecast errors tend to be larger for large events, which generally have larger impact than average events. Hence, this sensitivity is often desirable. As mentioned in section 3.4.4, the Nash–Sutcliffe efficiency (Nash and Sutcliffe, 1970), which is widely used in hydrology for model calibration, is a MSE skill score with the observed climatological mean as the reference forecast.

The components of the MSE under the CR decomposition (see Equation 62 in 0) may be estimated via (Murphy, 1973):

$$MSE = \frac{1}{n}\Sigma_{l=1}^{L}n_l(\overline{o_l} - f_l)^2 - \frac{1}{n}\Sigma_{l=1}^{L}n_l(\overline{o_l} - \overline{o})^2 + \frac{1}{n}\Sigma_{k=1}^{n}(o_k - \overline{o})^2 \tag{32}$$

where the first, second and third terms represent REL, –RES and UNC, respectively, n is the total number of forecast–observation pairs, L is the number of bins for the forecast events that the individual forecasts are subgrouped into, n_l is the number of pairs in the lth bin, $\overline{o_l}$ is the mean of the observations associated with the forecasts in the lth bin, f_l is the lth forecast event, $\overline{o}$ is the grand mean of all observations and o_k is the kth observation.

The components of the MSE under the LBR decomposition may be estimated via:

$$MSE = \frac{1}{n}\Sigma_{i=1}^{I}n_i\left(\overline{f_i} - o_i\right)^2 - \frac{1}{n}\Sigma_{i=1}^{I}n_i\left(\overline{f_i} - \overline{f}\right)^2 + \frac{1}{n}\Sigma_{k=1}^{n}\left(f_k - \overline{f}\right)^2 \tag{33}$$

where the first, second and third terms represent T2B, –DIS and SHA, respectively, n is the total number of forecast–observation pairs, I is the number of bins for the observed events that the individual observations are subgrouped into, n_i is the number of pairs in the ith bin, $\overline{f_i}$ is the mean of the forecasts associated with the observations in the ith bin, o_i is the ith observed event, $\overline{f}$ is the grand mean of all forecasts and f_k is the kth forecast.

The **root mean squared error (RMSE)** is the square root of the MSE.

The **mean error (ME)** is the average difference between forecasts and observations:

$$ME = \frac{1}{n}\Sigma_{k=1}^{n}(f_k - o_k) \tag{34}$$

The perfect score for the ME is 0. The ME is also known as the mean bias or first-order bias. A positive ME indicates a tendency to over-forecast, and a negative ME indicates a tendency to under-forecast. It is possible to have a small ME from large location- or time-specific forecast errors if the errors tend to cancel out with spatio-temporal averaging. Because mass balance is the most important governing principle in hydrology and water management, particularly at

the catchment scale, the ME or, equivalently, relative mean error (RME) (see below) or multiplicative bias (see below), is an extremely important measure in hydrological verification and should always be assessed for precipitation and streamflow over the range of spatio-temporal scales of mass-balancing interest.

The **relative mean error** (or relative bias) is the average difference between forecasts and observations relative to the observation mean:

$$RME = \frac{ME}{\mu_{obs}} \tag{35}$$

where

$$\mu_{obs} = \frac{1}{n}\sum_{k=1}^{n} o_k \tag{36}$$

The perfect score for the RME is 0. A positive RME indicates a tendency to over-forecast, and a negative RME indicates a tendency to under-forecast. RME has no units and is often expressed as a percentage instead of a fraction.

The **multiplicative bias** is a unitless measure of the overall (unconditional) bias defined as the ratio of the forecast mean to the observation mean:

$$\text{BIAS} = \frac{\sum_{k=1}^{n} f_k}{\sum_{k=1}^{n} o_k} \tag{37}$$

The perfect score for the multiplicative bias is 1. A value larger than 1 indicates a tendency to over-forecast, and a value smaller than 1 indicates a tendency to under-forecast. As with the ME, it is possible to score well on BIAS with significant location- or time-specific errors if they tend to cancel out when averaged in space or time.

The **Pearson correlation coefficient (CORR)** is a measure of the linear association between the forecasts and the observations:

$$CORR = \frac{\frac{1}{n}\sum_{k=1}^{n}\left[\left(f_k - \mu_{fcst}\right)\left(o_k - \mu_{obs}\right)\right]}{\sigma_{fcst}\sigma_{obs}} \tag{38}$$

where

$$\mu_{fcst} = \frac{1}{n}\sum_{k=1}^{n} f_k \tag{39}$$

$$\mu_{obs} = \frac{1}{n}\sum_{k=1}^{n} o_k \tag{40}$$

$$\sigma_{fcst} = \left[\frac{1}{n-1}\sum_{k=1}^{n}\left(f_k - \mu_{fcst}\right)^2\right]^{\frac{1}{2}} \tag{41}$$

$$\sigma_{obs} = \left[\frac{1}{n-1}\sum_{k=1}^{n}\left(o_k - \mu_{obs}\right)^2\right]^{\frac{1}{2}} \tag{42}$$

The score for CORR ranges from –1 to 1, and the perfect score is 1. Visually, correlation measures how tight the scatter is in any direction between the forecasts and the verifying

observations. A value of 0 indicates no correlation between the forecast and the observation. Correlation of –1 indicates perfect negative correlation, in which case changing the sign of the forecasts will produce perfect positive correlation. Correlation is immune to additive or multiplicative biases in the forecast and is sensitive to extreme values.

4.3.2 Scores for continuous probabilistic forecasts

The verification metric used most frequently to describe the accuracy of probabilistic forecasts is the mean continuous ranked probability score (CRPS) (Matheson and Winkler, 1976; Hersbach, 2000).

The **continuous ranked probability score (CRPS)** is a measure of the integrated squared difference between the cumulative distribution function of the forecast and the unit step function representation of the verifying observation. The CRPS for the *k*th forecast–observation pair is given by:

$$CRPS_k = \int_{-\infty}^{\infty} [F_k(x) - O_k(x)]^2 \, dx \tag{43}$$

where $F_k(x)$ is the forecast CDF and $O_k(x)$ is the unit step, or Heaviside, function representation of the verifying observation. The perfect score for the CRPS is 0 (when the forecast is perfectly sharp and equals the observation).

The CRPS for a given forecast–observed pair may be visualized by plotting the CDF of the forecast and the unit step (Heaviside) function representation of the observation as shown in Figure 16. The CRPS corresponds to the area between the two functions. In practice, the CDF is usually approximated by the empirical CDF (the "staircase" CDF) defined by the ensemble members.

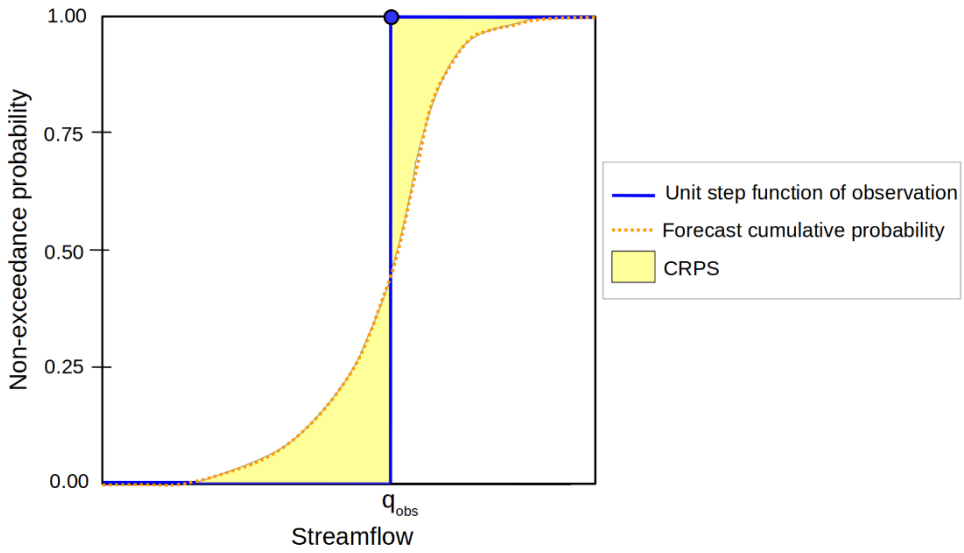


Figure 16. Illustrative example of the CRPS for a given pair of a probabilistic forecast and the verifying observed flow value q_{obs}

The mean CRPS is the average of all $CRPS_k$, $k = 1,..,n$:

$$\overline{CRPS} = \frac{1}{n} \Sigma_{k=1}^{n} CRPS_k \tag{44}$$

The CRPS extends the RPS with an infinite number of categories (Hersbach, 2000) and is equivalent to the BS integrated over all possible thresholds within the range of the forecast (Jolliffe and Stephenson, 2012; Gneiting et al., 2005; Hersbach, 2000). Unlike the RPS, the

CRPS has the same unit as the predictand, which facilitates the assessment of the overall accuracy of probabilistic forecasts in physical space. One may hence consider the mean CRPS as a representative measure of accuracy for probabilistic forecasts, as the RMSE is for that of single-valued forecasts. The CRPS skill score (CRPSS) is defined according to Equation 2.

The mean CRPS for single-valued forecasts (that is, perfectly sharp ensemble forecasts) reduces to the MAE, which allows possible comparison between probabilistic and single-valued forecasts. It is important to recognize, however, that such a comparison necessarily favours probabilistic forecasts, which reduce the risk of being wrong by spreading possible outcomes, over single-valued forecasts, which bet everything on a single outcome.

Hersbach (2000) describes the CR decomposition of the mean CRPS:

$$\text{Mean CRPS} = UNC_{CRPS} + REL_{CRPS} - RES_{CRPS} = REL_{CRPS} + CRPS_{pot} \quad (45)$$

where $CRPS_{pot}$ is the potential mean CRPS representing the mean CRPS one could potentially obtain when the forecasts are perfectly reliable or well-calibrated (that is, $REL_{CRPS} = 0$). The mean CRPS, as well as the reliability component and the potential mean CRPS are all negatively oriented, with a perfect score of 0. Unlike the reliability component of the BS, REL_{CRPS} measures reliability similarly to the rank histogram (see section 0), that is, it assesses if the verifying observation is statistically just another ensemble member of the forecast distribution. The potential mean CRPS, $CRPS_{pot}$, is sensitive to the average ensemble spread and the frequency and magnitude of outliers. For the best $CRPS_{pot}$, the forecasting system needs a narrow ensemble spread on average without ensemble outliers that are too numerous or too high (Hersbach, 2000).

4.3.3 Widely used diagrams for verification of probabilistic forecast

This section describes the reliability diagram, the Talagrand diagram (or rank histogram) and the probability integral transform (PIT) histogram used for assessing reliability, and the relative operating characteristic (ROC) curve and discrimination diagram used for assessing discrimination.

The **reliability diagram** plots the forecast probability of an event of interest, or simply forecast probability, on the x-axis and the fraction of the times when the event is observed given the forecast, or observed relative frequency, on the y-axis. If the forecast is reliable, the two should agree. For example, for all forecasts that predicted an event with a 25% chance, the event should have occurred 25% of the time whenever the forecast probability was 0.25. Each reliability diagram is constructed with respect to a specific binary event of interest (for example, the observed flow or some other variable exceeds some threshold). Hence, multiple reliability diagrams are generally necessary to assess the reliability of, for example, ensemble streamflow forecast across a range of magnitudes of streamflow.

Because reliability is a conditional measure (see Equation 59 in Appendix B), it is necessary to divide the forecasts into bins of varying levels of probability. Hence, the reliability diagram requires a large dataset so that the sample size is adequate for each bin, particularly for events associated with high probability thresholds such as flooding. For the diagram's construction, the forecast probability is divided into K bins between 0 and 1 so that the sample size is not too small for any bin. Typically, the reliability diagram includes the histogram of the sample size of forecast probability across all bins to assess sharpness.

When the lines of no resolution and no skill (in reference to climatology) are added, the reliability diagram is called the attributes diagram (Hsu and Murphy, 1986), as shown in Figure 17. In the reliability diagram part of Figure 17, the forecast probabilities are binned into 10 subranges of equal interval. For each subrange of forecast probability, the associated solid red circle indicates the observed relative frequency on the y-axis (that is, the fraction of the verifying observations of the event among all forecasts that predicted the event with the probability shown on the x-axis).

In Figure 17, the horizontal no-resolution line corresponds to the climatological mean of the observation (in probability space). Recall that climatological forecasts are reliable but have no resolution. The no-skill line, for which climatology is the reference forecast, is given by the midpoints between the no-resolution line and the diagonal or perfect-reliability line (Hsu and Murphy, 1986). The shaded area in the attributes diagram is where the forecast contributes positively to the overall skill in reference to the climatological forecast.

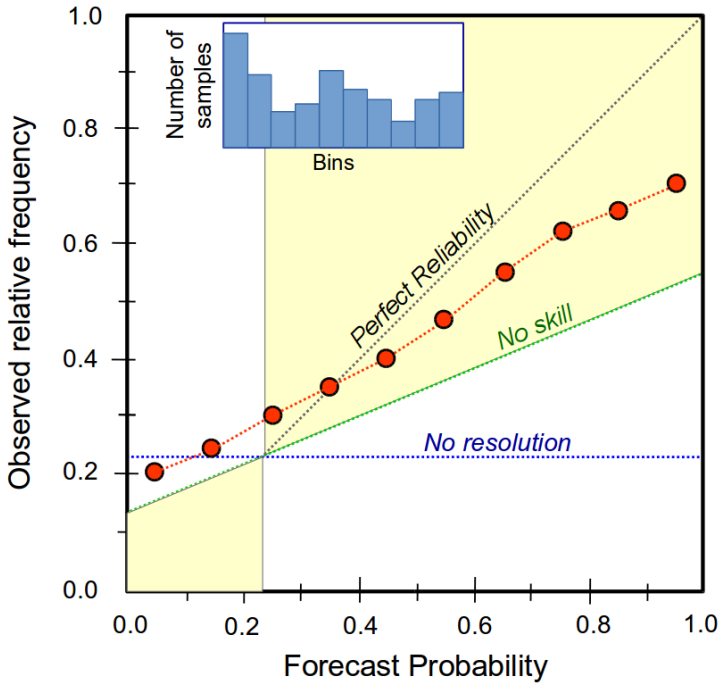


Figure 17. Illustrative example of the attributes diagram

Perfectly reliable forecasts have the points representing forecast probability and observed relative frequency (the reliability curve) along the diagonal in the reliability diagram. Significant departures from the diagonal indicate type I conditional bias. Figure 18 provides several illustrative examples of the reliability curve. A reliability curve that lies above the diagonal is indicative of under-forecasting (in probability) – that is, the observed relative frequency of the event is greater than the forecast probability. A reliability curve that lies below the diagonal is indicative of over-forecasting.

Forecasts are under-confident (that is, the ensemble members are over-spread) if the event occurs more frequently than the forecast probability. Forecasts are over-confident (that is, the ensemble members are under-spread) if the event occurs less frequently than the forecast probability. A horizontal reliability curve indicates no resolution, meaning that events occur with the same frequency regardless of the forecast probability. A negatively sloped reliability curve indicates negative reliability (that is, negative type I conditional bias), meaning that events occur more frequently when the forecast probability is smaller and less frequently when it is larger.

If the sample size is small, not all bins for forecast probability may be represented, leading to interruptions in the reliability curve and hence an erratic pattern due to the sampling uncertainty. In this case, one may reduce the number of bins so that the sample size increases for the data-deficient bin(s). If the above does not produce stable reliability diagrams, other reliability measures may be used, such as the reliability component of the BS and that of the mean CRPS (see section 4.3.2). It is important to recognize, however, that the above reliability components are weighted averages over all forecast probabilities and hence are not as strong a test for reliability as the reliability diagram for specific events. In addition, unlike the reliability component of the BS, the reliability component of the mean CRPS assesses the likeness of the probability distribution of the ensemble members to that of the verifying observation, similarly to the rank histogram (see below).

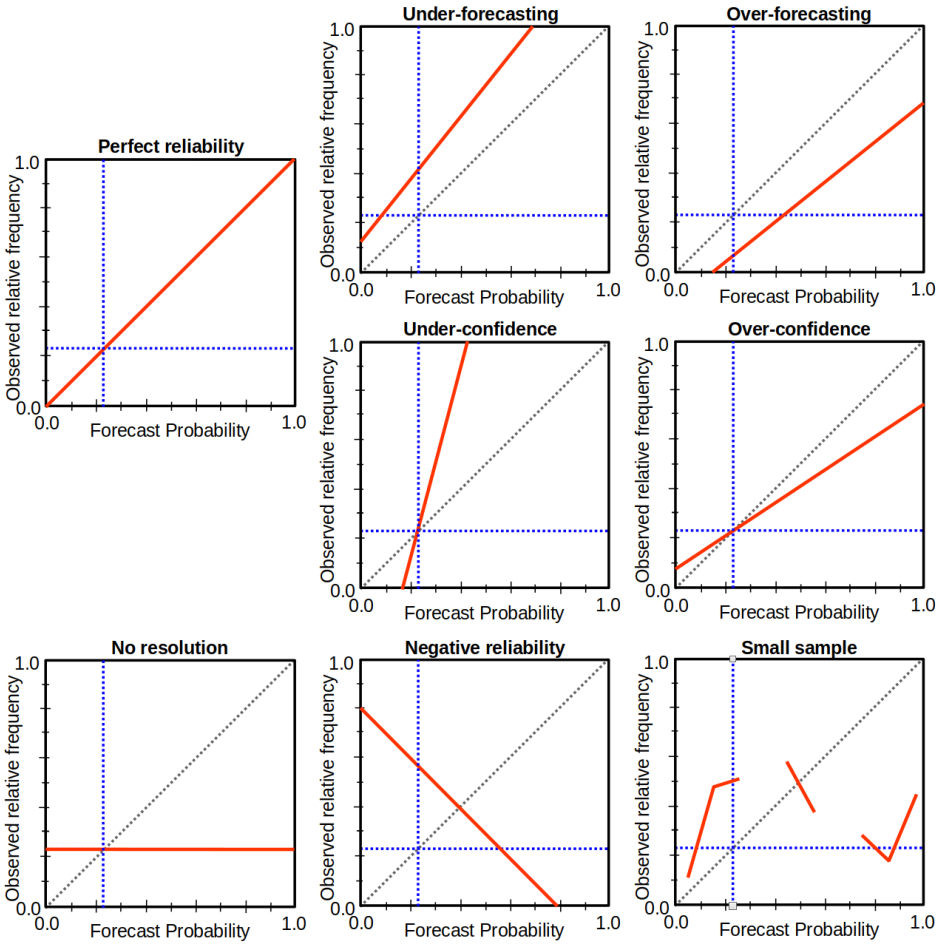


Figure 18. Illustrative examples of the reliability diagram with interpretation for a defined event of interest

Reliability of ensemble forecasts may be assessed with the Talagrand diagram, also known as the rank histogram; however, this is weak test of reliability. Ensemble forecasting aims at each ensemble member statistically behaving like the observation. The Talagrand diagram assesses if the ensemble spread is consistent with the variability of the observation. A uniform Talagrand diagram is hence a necessary, but not sufficient, condition for reliability (in the sense of lack of type I conditional bias) and is often used as an initial check on reliability (Hamill 2001).

The **Talagrand diagram (**or **rank histogram)** measures how well the ensemble spread represents the variability of the observations by plotting the fraction of the observations that fall between any two ranked ensemble members. The plot verifies if the observation is equally likely to occur in each of the $k + 1$ bins of an ensemble forecast with k members. The Talagrand diagram is created via the following steps: order the k members in an ensemble forecast from the smallest to the largest, define the $k + 1$ bins between any two ranked ensemble members and identify the bin that the observation falls into, and repeat the above for all forecast–observation pairs while counting how many observations fall into each bin. Since the bins are defined for each ensemble forecast based on ranking the members, they vary in size. Also, there is a probability of $2 / k$ that the observation will fall outside of the ensemble spread. For a uniform rank histogram, the number of observations in each bin is $N / (k + 1)$, where N is the total number of forecast–observation pairs.

Figure 19 presents illustrative examples of the rank histogram and what they may indicate. If the ensemble members perfectly represent the variability of the observation, an observation is equally likely to fall into any bin, resulting in a flat Talagrand diagram showing rank uniformity. An asymmetric histogram usually indicates biases in the ensemble mean or possibly in higher-order moments. An L-shaped Talagrand diagram, which reflects greater observed frequencies

towards the lower categories, indicates that observations are too often smaller than the ensemble members, usually owing to over-forecasting. If the Talagrand diagram shows greater observed frequencies towards the higher categories, observations are too often larger than the ensemble members, usually owing to under-forecasting. A symmetric, dome-shaped Talagrand diagram indicates that not enough observations are falling at the extremes, as most observations fall near the centre of the ensemble spread. In such a case, the ensemble spread may be too large, owing to over-dispersion. A U-shaped Talagrand diagram usually indicates that too many observations fall at the extremes of the ensemble spread, usually owing to under-dispersion.

Talagrand (1997) proposed a metric based on the deviation from a flat histogram to characterize the forecast reliability in the context of rank histogram and to compare different sets of ensemble forecasts. Hamill (2001) reports that a U-shaped diagram, commonly interpreted as indicating a lack of ensemble spread, may also be a sign of conditional bias. Also, when using imperfect observations, the observational error could impact the shape of the rank histogram. Typically, the larger the observational error is, the more U-shaped the rank histogram will appear, even for reliable ensemble forecasts. Therefore, lack of rank uniformity should be further investigated with other verification measures and diagrams to ascertain the nature of the forecast errors and avoid misinterpretation.

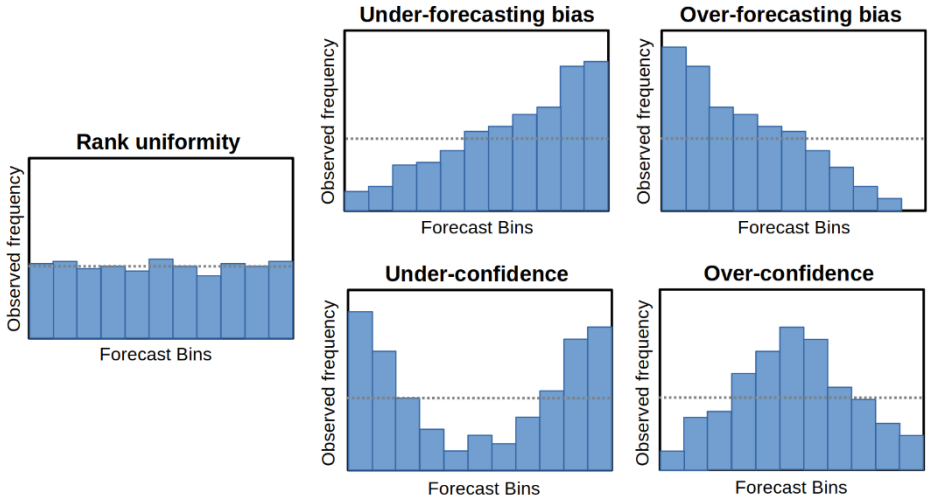


Figure 19. Illustrative examples of the Talagrand diagram and their interpretation

The **probability integral transform (PIT) histogram** is an analogue of the rank histogram and has similar interpretations. Reliable probabilistic forecasts yield PIT histograms that are flat or uniform, whereas U- and dome-shaped PIT histograms are often associated with under- and over-dispersed forecasts, respectively. The PIT diagram is the cumulative distribution of the PIT values (Gneiting et al., 2007; Wilks, 2011). For reliable forecasts, the PIT diagram falls on the 45° line. Similarly to the interpretation of the rank histogram, deviations from the diagonal line in the PIT diagram may help diagnose forecast errors. Case 6 in Chapter 7 uses the PIT histogram to assess reliability of ensemble streamflow forecast for ephemeral streams.

For discrimination, the most widely used diagram is the ROC curve.

The **relative operating characteristic (ROC)** curve is a signal detection curve for binary forecasts and plots POD (or hit rate) on the y-axis versus POFD on the x-axis (see section 4.2 for definitions of POD and POFD). Developed originally in electrical (radar) engineering (Green and Swets, 1966), the ROC curve measures discrimination by conditioning on the observations (that is, on the observed yes-events for POD and the observed no-events for POFD).

For perfect discrimination, the curve travels from (0,0) to (0,1) at the top left of the diagram (POFD = 0 and POD = 1), then to (1,1) at the top right of the diagram. The diagonal line indicates no skill (POD = POFD) of a random forecast (that is, a coin flip). The forecast has

discriminatory skill if its ROC curve is above the diagonal line. A ROC curve below the diagonal indicates negative discriminatory skill (the observed outcomes of the forecast no- and yes-events are likely to be yes- and no-events, respectively). For single-valued forecasts (that is, perfectly sharp probabilistic forecasts), the ROC curve forms a triangle connecting (0,0), (POFD, POD) and (1,1).

Figure 20 illustrates the ROC curve and the area under the curve (AUC) for prediction of flooding or no flooding using single-valued forecasts. In Figure 20., the blue solid lines represent the ROC curve, the ROC score is given by the AUC, the diagonal (POD = POFD and AUC = 0.5) represents the line of no discriminatory skill, and the ROC curve below the diagonal represents negative discriminatory skill. The closer POD and POFD are to 1 and 0, respectively, the larger the discriminatory skill is. The shape of the ROC curve is triangular in this illustrative example because single-valued forecasts are equivalent to single-member ensemble forecasts and hence perfectly sharp. Therefore, the forecast probability of exceedance with respect to the threshold of interest is either 0 or 1.

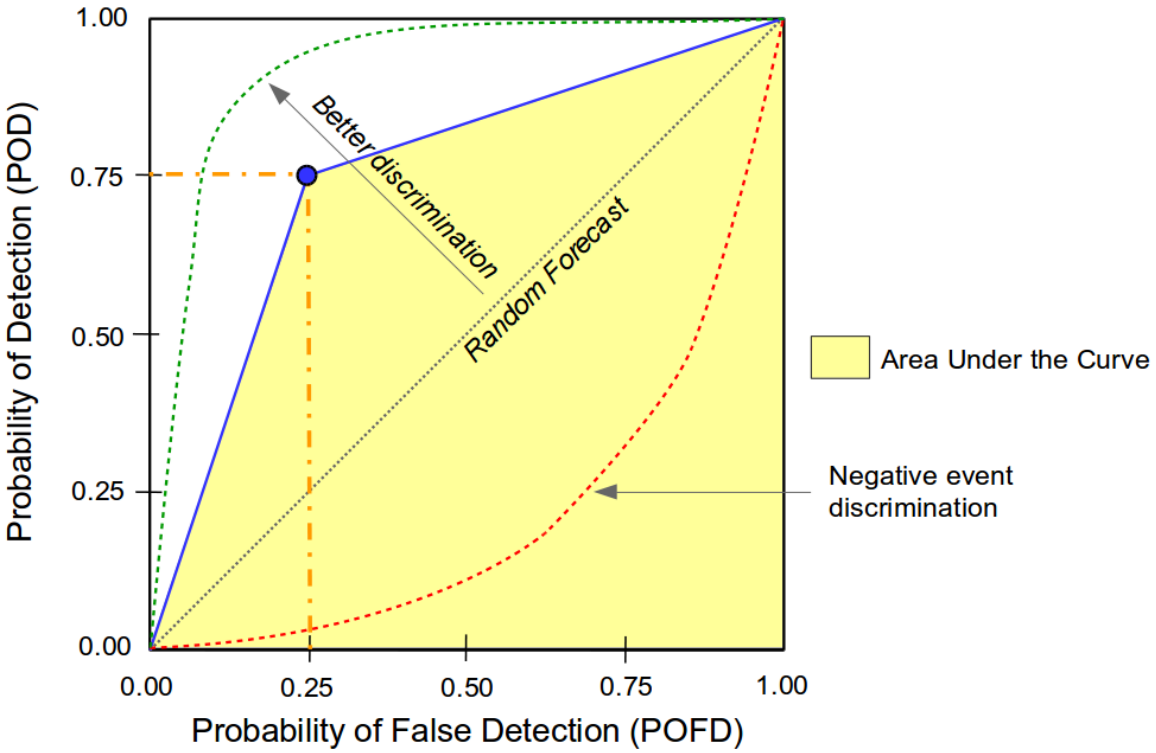


Figure 20. ROC curve (in blue) for single-valued, or perfectly sharp, binary forecasts with POD of 0.75 and POFD of 0.25

For probabilistic forecasts, the ROC curve is obtained by plotting POD versus POFD using a set of increasing probability thresholds to define the forecast yes- and no-events. For each of the *N* increasing probability thresholds, a forecast is counted as a yes-event if the forecast probability is above the probability threshold, and as a no-event otherwise. For each probability threshold, the POD and POFD scores are computed from the 2 × 2 contingency table (see section 4.2). The ROC curve for probabilistic forecasts then connects (0,0), ($POFD_j$, POD_j) for j = 1 to N, and (1,1).

Figure 21 illustrates the ROC curve and the AUC for discrimination of flooding versus no flooding for probabilistic forecasts. In this example, the decision probability thresholds used are 0.2, 0.4, 0.6 and 0.8. Figure 21 shows that the probability threshold of 0.4 yielded *PODj* of 0.75 and *POFDj* of 0.3 for the probability forecasts. It is readily seen in Figure 21 how one might choose a personalized or application-specific probability threshold to reflect one's aversion or tolerance for POFD.

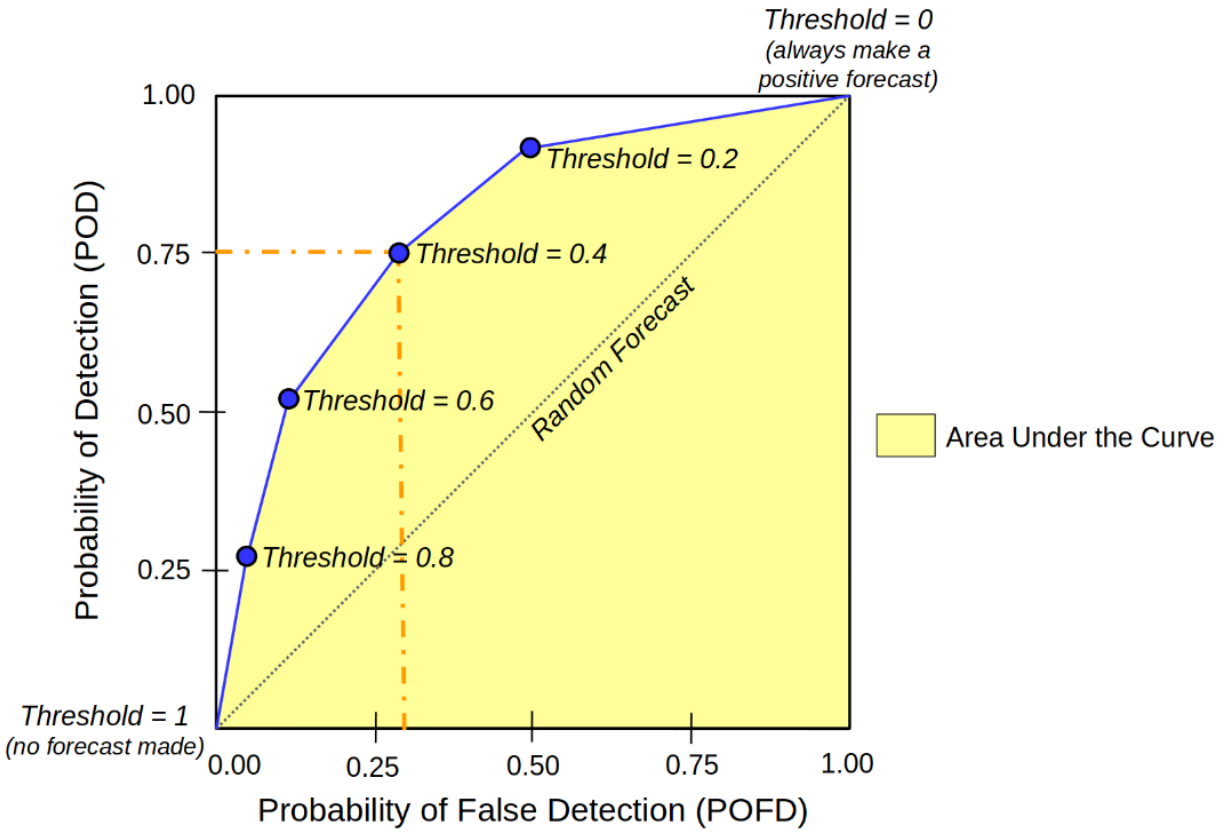


Figure 21. ROC curve (in blue) for probabilistic binary forecasts for which POD_j and $POFD_j$ are evaluated for a set of j increasing decision probability thresholds of 0.2, 0.4, 0.6 and 0.8

Nominally, ROC curves for a single-valued forecast and a probabilistic forecast based on the same event definition may be intercompared if plotted together. Such a comparison, however, necessarily favours a probabilistic forecast which adds curvature to the ROC curve by hedging the bets (spreading the possible outcomes around). As seen in Figure 20, the ROC "curve" for single-valued forecasts can only be triangular by betting all on a single outcome and hence has a smaller AUC. A more useful comparison, particularly if the underlying hydrological models used are the same, is to check the location of the (POFD, POD) for the single-valued forecast relative to the AUC for the probabilistic forecast. If the former lies outside of the latter (that is, farther away from the diagonal than the farthest point in the ROC curve from the diagonal), it is an indication that one may be able to improve the quality of probabilistic forecast by improving uncertainty modelling.

The **ROC score** is a summary score for the ROC curve and is defined as:

$$ROC\ Score = 2 \times (AUC - 0.5) \tag{46}$$

The ROC score is 0 (that is, no discriminatory skill) if the ROC curve corresponds to the diagonal (POD = POFD). The perfect score is 1 (POD = 1 and POFD = 0).

Discrimination may also be visually assessed with the discrimination diagram.

The **discrimination diagram** plots, for each of all mutually exclusive and collectively exhaustive verifying events, such as flooding and no flooding, the conditional probability that the event is forecast given that the event is observed (on the y-axis) as a function of the (unconditional) probability of the event being forecast (on the x-axis). For binary events, this diagram plots the conditional probability that the yes-event was forecast given that the event occurred, and the conditional probability that the no-event was forecast given that the event did not occur as functions of the probability of the yes-event being forecast. If the forecasts

have good discriminatory skill, the resulting two distributions, or so-called likelihood functions, would be well separated from each other, exhibiting two distinct peaks with little overlap.

As may be inferred from the above, the discrimination diagram is similar to the ROC curve for binary events. The main difference is that, whereas the latter plots POD versus POFD for a range of different exceedance probabilities of the yes-event being forecast, the former plots POD and 1 – POFD for varying levels of the probability of the yes-event being forecast. Figure 22 illustrates the discrimination diagram for probabilistic forecasts for binary events such as flooding and no flooding. The discrimination diagram consists of two plots: the fractions of the yes-events forecast among all yes-events observed (in red) and the fractions of the no-event forecasts among all no-events observed (in blue).

All fractions mentioned above are calculated for and connected across the 5 different levels of probability of yes-event forecast shown on the x-axis. The probability levels are chosen so that the sample size is adequate for each bin. The resulting two likelihood functions exhibit some separation, showing the forecasts' ability to differentiate yes- and no-events. The lower plots show discrimination diagrams for forecasts with good discrimination (left) and forecasts with poor discrimination (right). To characterize the separation between the two likelihood functions, one may calculate the discrimination distance defined as the absolute difference in mean between the two likelihood functions.

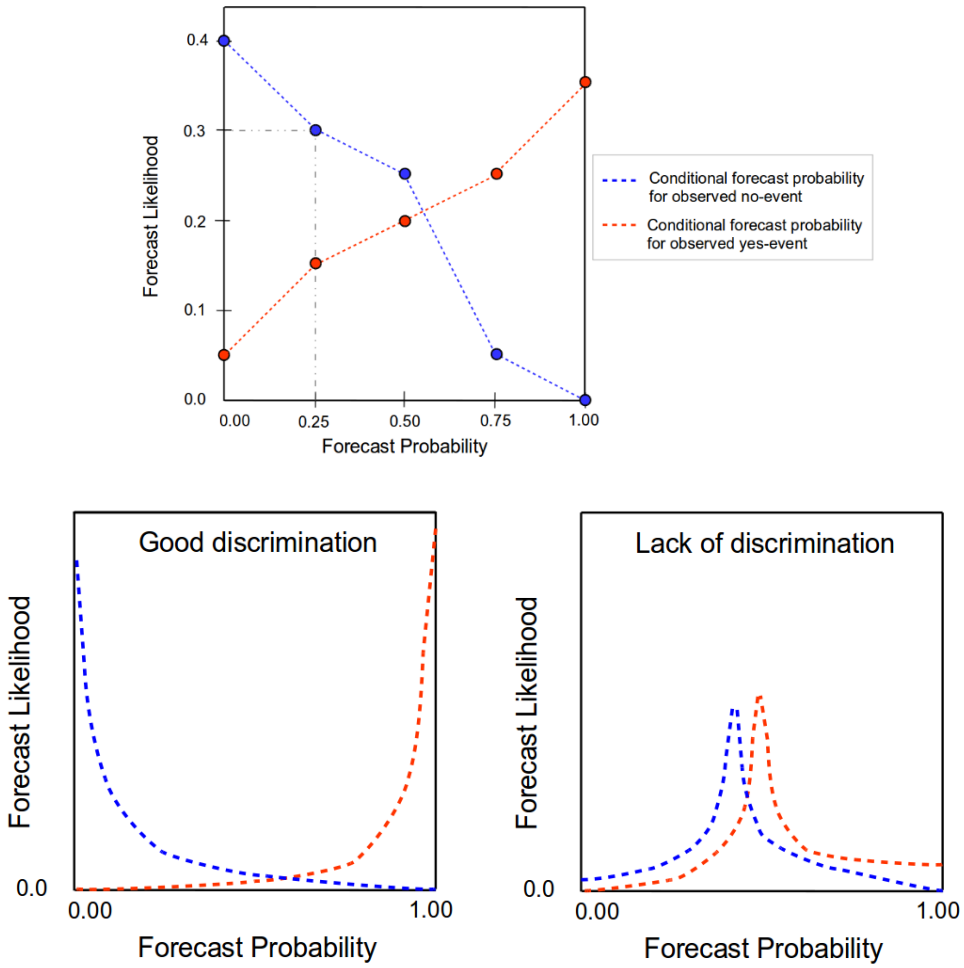


Figure 22. Illustrative examples of discrimination diagram for a binary probabilistic forecast

4.4 Additional considerations

As forecast accuracy depends on various attributes, several verification metrics and diagrams are typically necessary to adequately assess the quality, strengths and weaknesses of the forecast. The choice of the most suitable metrics for the verification task at hand should be based on a step-by-step approach, starting with the selection of at least one metric for each forecast attribute of interest. The choice of verification metrics depends not only on the application but also on the verification objectives, as illustrated in the case studies presented in Chapter 7 and in the examples from the verification literature reviewed by Anctil and Ramos (2019).

In hydrological model calibration, one may employ a set of performance evaluation criteria to assess the quality of model simulation. For example, Moriasi et al. (2015) reports ranges of values for performance criteria for a small number of metrics selected at the catchment and field scales and rates the performance from "very good" to "not satisfactory". For verification of streamflow forecasts, however, such an approach is not likely to yield verification information that is very useful for user decision-making for multiple reasons, as explained below. Whereas the accuracy of model-simulated streamflow reflects only the structural and parametric uncertainties in the hydrological model, the accuracy of streamflow forecasts depends additionally on the accuracy of the input forecast, the memory of the hydrological system (Alizadeh et al., 2020) and the collective predictive skill of the forecast system, which reflects not only model calibration but also DA, postprocessing and possibly other factors (see section 2.4). Because the above predictability and predictive skill depend not only on the lead time but also on the temporal scale of aggregation and the magnitude of streamflow, assessing the predictive skill of a forecast with simulation-centric performance evaluation criteria is likely to be too limiting. On the other hand, an operational agency may select a small number of lead time- and scale-specific verification metrics to measure and track the improvement over time in the quality of river forecasts.

As verification analysis is often used to compare forecast quality for different forecast locations, it is important to select metrics and thresholds such that the verification scores may be compared across different locations. To intercompare or aggregate verification results at different locations, normalized verification metrics such as skill scores may be necessary. Skill scores measure percent improvement (or deterioration) in the metrics of choice by the subject forecast over the reference forecast of choice. They also help determine if the verification results are good because the forecast system has good predictive skill or because the predictand is very predictable.

Depending on the application and the choice of the reference forecast, skill scores may not, by themselves, inform the user of the practical significance of the improvement for decision-making. For example, a 10% increase in MSESS or CRPSS of the subject forecast in reference to climatology for the dry season is likely to carry far less significance for flood forecasting than that for the wet season. For ensemble water quality forecasting (Kim et al., 2014), on the other hand, such an increase in skill score in the dry season is likely to be significant. Recall also in section 4.2.2 the counterintuitive illustrative example involving the BSS.

It is good practice to examine the verification metrics themselves in addition to the skill scores to gauge hydrological significance and potential impact to the user's decisions. Many users of forecast information do not have the resources or expertise to translate verification information through their decision support systems or processes to assess application-specific impact. For this reason, the RMSE and mean CRPS are particularly useful (and hence popular) because they are expressed in the same unit as the predictand itself. For example, a dam operator will be able to relate the reductions in the RMSE or mean CRPS in units of discharge and volume with the reservoir's release and storage capacities far more easily than any skill score.

For those verification metrics that require binary representations of the forecast and the verifying observation such as the reliability diagram and the ROC, the event categorization (the prime example being flooding and no flooding) should be made commonly across all

metrics to allow cross-assessment of different attributes in the same reference frame. The event categorization may be for flooding and no flooding, surface runoff and no surface runoff, multiday inflow volume exceeding and not exceeding some threshold, and so on. Regardless of the choice, the categorized events must be mutually exclusive and collectively exhaustive. To facilitate intercomparison among different locations, the threshold used in the categorization may be defined as percentiles in the observed distribution (for example, 10th percentile for low flow and all other flows, or 90th percentile for significant flow and all other flows). When using common event categorization to assess forecast quality across multiple basins, the verification results may be more difficult to interpret due to the likely variations in sample size, particularly for large-to-extreme events. Preferably, one should assess sampling uncertainty by estimating confidence intervals analytically or numerically via bootstrapping. At minimum, one should communicate sampling uncertainty by reporting the number of forecast–observation pairs as a proxy (in the entire dataset as well as for the subsamples when applying a specific condition or conditions). Verification plots with confidence intervals are particularly useful to users who can integrate the information for improved decisions.

The adequacy of the sample size may be assessed by calculating confidence intervals for the verification metrics of choice (see Figure 47 for an example). If the confidence interval is acceptably narrow based on significance testing or visual examination of the width of the interval relative to that of the verification metric itself, one may consider the sample size adequate. In practice, such assessment is often not very straightforward, as it may require examination of many confidence interval-assessed verification metrics following stratification or conditioning of the data with respect to duration, headwaters versus downstream, season, magnitude of flow and possibly other hydroclimatological attributes. Note that, depending on the extent of the stratification or conditioning, the sample size is likely to be greatly reduced. The above examination is also necessary to ascertain the magnitude of the largest events that may be verified with statistical significance given the period of record. To increase sample size, one may forego stratification or conditioning at the expense of losing specificity with respect to certain hydroclimatological attributes or lowering the magnitude of the largest events being verified. Often, assessment of such trade-offs is not clear cut, and judgment will have to be exercised based on experience and practicality. If location-specific verification is not possible, one will have to trade location specificity for increased sample size via aggregate verification (see Case 2 in Chapter 7). The guidance on sample size obtained from the above process may be transferable to other locations of similar hydroclimatology and period of record. In such cases, repeated assessment of confidence intervals based on, for example, bootstrapping (see section 2.4), which is computationally expensive, may be avoided.

Often, verification tasks involve comparing forecasts from two competing forecast systems or processes using identical verifying observations. When evaluating confidence intervals for the differences in verification results in such head-to-head comparative verification, care should be taken to account for the likely positive correlation between the competing forecasts. Estimating confidence intervals independently of each other when the two forecasts are significantly correlated (that is, statistically dependent) is likely to overestimate the sampling uncertainty associated with the difference in the metric of interest calculated for the two forecasts. For example, when comparatively verifying forecast A with forecast B versus common verifying observations using mean CRPS, the sampling uncertainty of interest may be for $\Delta\overline{CRPS} = \overline{CRPS}_A - \overline{CRPS}_B$ rather than for $\overline{CRPS}_A$ and $\overline{CRPS}_B$ individually. If the lower uncertainty bound for $\Delta\overline{CRPS}$ is positive, one may conclude that the reduction in mean CRPS by forecast B over forecast A is statistically significant. If bootstrapping is used to assess the sampling uncertainty of $\Delta\overline{CRPS}$, identical realizations of random sampling with replacement should be used between the two forecasts when calculating $\overline{CRPS}_A$ and $\overline{CRPS}_B$.

In operational hydrological forecasting, accurate prediction of large-to-extreme events is often the most important service. In the descriptions thus far, it is implicitly assumed that, once the forecast–observation pairs are stratified as necessary with respect to season (for example, cool versus warm or dry versus wet) and possibly other hydroclimatological or physiographic attributes, one may obtain all necessary verification information by carrying out verification using all available forecast–observation pairs. Recall in Chapter 3 that the distributions-

oriented approach assumes the forecast–observation pairs to be jointly IID. In reality, the forecast–observation pairs within the periods of significant hydrograph response are likely of different statistical character than those in the rest of the periods. Similarly, the pairs from different hydrological or hydraulic regimes (for example, surface runoff and interflow versus baseflow, in-channel versus out-of-channel flow) may not share the same joint probability distribution. To assess such flow regime-dependent forecast quality, it is often desirable to perform verification conditionally on certain events in addition to "unconditional" verification. Conditional verification amounts to decomposing the verification metrics of choice into regime-specific contributions to help address the above questions, as illustrated below using the MSE and low versus high flow as an example:

$$MSE = MSE_{low\ flow} \cdot fr_{low\ flow} + MSE_{high\ flow} \cdot \left(1 - fr_{low\ flow}\right) \quad (47)$$

where $MSE_{low\ flow}$ and $MSE_{high\ flow}$ are the MSEs for the low- and high-flow periods, respectively, and $fr_{low\ flow}$ and $\left(1 - fr_{low\ flow}\right)$ are the fractions of the low- and high-flow periods, respectively. In the above, the conditioning may be based on any events of choice as long as they are mutually exclusive and collectively exhaustive so that Equation 47 is a proper identity in the mean sense.

The additional verification information from such analysis often provides additional insight into diagnosing the performance of forecast systems and processes, and aids regime-specific characterization of forecast quality. Cases 1 through 4 in Chapter 7 provide examples of using conditional verification for such purposes. It is important to note that, in general, conditional verification results should be accompanied by the "parent" unconditional verification results to avoid misinterpretation or misuse. For example, using only the conditional verification information amounts to using improper scoring (Bellier et al., 2017; Lerch et al., 2017) and will inevitably lead to poor decisions most of the time given that low flows occur far more frequently than high flows.

Currently, there is no settled way to address the independence assumption in streamflow verification. Baseflow tends to be highly correlated in time, whereas "event" flow is much less so, depending on the sampling interval. In addition, the collective baseflow period is generally significantly larger than the collective event flow period (that is, $fr_{low\ flow} > fr_{high\ flow}$ in Equation 47). Hence, using all available forecast–observation pairs amounts to oversampling low-flow conditions. It is technically possible to assess regime-dependent temporal correlation and subsample the pairs accordingly. Such statistical modelling, however, is subject to errors of its own due to possible sampling uncertainty, skewness, heteroscedasticity and nonstationarity. In addition, such subsampling will render metrics such as the ME – which is critical to assessing mass balance – difficult to interpret. To reduce temporal correlation, one may aggregate the pairs over sufficiently large subperiods and verify the time-aggregated pairs. Such an approach, however, does not provide verification information for instantaneous flow, which is the most important piece of data for flood forecasting. The combined use of unconditional and conditional verification represents a practical compromise to produce verification information necessary for user decision-making without introducing additional layers of complexity. As hydrological verification develops further, better approaches are likely to emerge.

When verifying skewed and heteroscedastic variables such as streamflow (as well as precipitation), one may find that both type I and type II conditional biases may be acceptably small when all ranges of flow are considered, but that type II conditional bias is unacceptably large for high flows. Depending on the service goals of the forecasting agency, such additional verification information may potentially impact decisions about the choices of the input forecasts, hydrological models, forecast system components and forecast processes. For example, if the agency's most important service is flood forecasting, it may be willing to trade, up to a point, overall accuracy and reliability (that is, type I error) for significantly reduced type II error to improve predictive skill for large-to-extreme events. In the context of categorical forecasting, the above trade-off amounts to accepting more frequent false alarms for smaller events in favour of avoiding "big" misses.

To support such decision-making, proper verification is critically important, as one must be able to identify and assess the competing attributes and their trade-offs objectively and coherently, often in a zero-sum game. Accuracy measures such as the weighted version of the CRPS (Gneiting and Ranjan, 2011) attempt to incorporate magnitude-dependent importance into proper metrics by placing larger weights over a specific region of the predictand. It is unclear, however, whether such approaches are any more practical than providing the user with additional "predictand region-specific" verification information so that one may make one's own decisions. Practicing hydrologists and water resources engineers are well prepared to use such information because they routinely communicate in terms of return periods or frequencies of occurrence (as in flood frequency analysis) and hence clearly understand what conditioning on the verifying observation exceeding some percentile represents versus no such conditioning.

An important group of users of hydrological forecasts and hence verification information are the operators and managers of infrastructure for control, conveyance and mitigation of floods, including urban drainage systems, and those of other critical infrastructure. The design storms and discharges for such systems are typically expressed in terms of the return period (Chow et al., 1988). Hence, these users are particularly interested in forecast skill for precipitation or discharge events exceeding certain thresholds. Societal needs for such additional verification information are likely to grow – particularly for flooding in urban and suburban areas as urbanization continues globally in addition to climate change (Ritchie et al., 2024). Such changes mean increasing nonstationarity in the hydrometeorological and hydrological processes (Milly et al., 2008), which makes calibration for reliability increasingly challenging while increasing the chances of "surprises" (that is, misses). Hydrological verification must be able to support decision-making in such an environment. The above picture underscores the importance of increasing the sample size so that one is able to make stronger inferences about the right tail of the distribution. To help reduce such observational information gaps, it may be necessary to consider non-traditional forms of observation as well in hydrological verification (Noh et al., 2019).

4.5 Key points

- Multiple metrics are available to assess each of the widely measured attributes for different types of forecasts. The MSE (or RMSE), the 2 × 2 contingency table, the BS and the mean CRPS are particularly important, as they collectively contain almost all other metrics or their building blocks. It is hence important to understand what each of the above four metrics comprise, so that one may utilize the entire suite of metrics, scores and diagrams effectively.

- This publication describes most of the widely used verification metrics, scores and diagrams. Some of the metrics derived from the 2 × 2 contingency table are better suited for verification of flash flood forecasts than streamflow forecasts. All others described in this chapter apply to the verification of streamflow forecasts. Most of them are available in verification software tools such as the EVS.

- Several diagrams and histograms are used to assess the attributes associated with the BS decomposition for probabilistic forecasts. The most widely used include the reliability diagram (a strong test of reliability), the rank histogram (a weak test of reliability), the ROC curve for discrimination and the forecast frequency histogram for sharpness. Depending on the application, the PIT diagram (see Case 6 in Chapter 7) and discrimination diagram may be preferred to the rank histogram and the ROC curve, respectively.

- Streamflow and precipitation typically exhibit skewness (asymmetry in distribution) and heteroscedasticity (nonuniformity in variability), which are often not explicitly assessed in verification. For streamflow, skewness and heteroscedasticity reflect predictability and flow regime-dependent variability, respectively, and hence provide very useful guidance on stratification, pooling, or conditioning of the forecast–observation pairs.

- Sampling uncertainty is a recurring challenge in hydrological verification, particularly for conditional verification of probabilistic forecasts for large-to-extreme events. It is a good practice to assess sampling uncertainty via bootstrapping early in the verification task for sample size-challenged cases. One may then assess how sampling uncertainty may be reduced by relaxing the conditioning or trading space for time via data pooling or regionalization.

- Conditional verification results, if produced, should be communicated together with the "parent" unconditional verification results to avoid misinterpretation or misuse. For example, using verification information for high flows as being representative of all flows will inevitably lead to poor decisions most of the time.

CHAPTER 5. PREPARATORY STEPS AND LOGISTICAL CONSIDERATIONS

The ultimate goal of verification is to support and improve the decision-making of users of the forecast products and services. Whether it is implicit or explicit, there is often a process in place in which the verification informs a decision which in turn triggers a response. This chapter concerns the production of verification information (the left-most process in Figure 23).

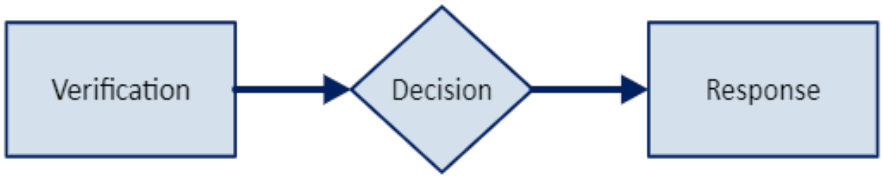


Figure 23. Verification, decision, response process

In practice, several steps are generally necessary to produce useful verification results. Each step involves multiple choices and decisions which, if not well thought out in advance, may lead to avoidable trials and errors. This chapter provides the reader with practical points to consider for purposeful and time-effective verification.

5.1 Defining verification objectives

Verification tasks should be planned and designed such that objectives are clear and well defined. Several example tasks of widely varying scope and specificity are presented below based on their main objective.

Example I: Provide the users of forecast information with basic forecast accuracy information. From the perspective of an operational forecasting agency, arguably the most important objective for verification is to provide users of forecast information with an up-to-date summary of the past performance of their forecast products. Whereas it is generally impractical to tailor the verification information to specific user groups, it should be reasonably location-specific and reflective of seasonality and other broadly important hydroclimatological and physiographic attributes. Such information is likely to aid the users' decision-making tangibly and hence increase the utility and value of the agency's forecast products. For this objective, the verification task should broadly address the basic user question of "How good is your forecast?"

Example II: Provide human forecasters and users with organization-specific risk profiles with additional forecast accuracy information. For potential large-impact events, managers of emergency response, public safety, critical infrastructure (including that for flood control and transportation), and human, material and natural resources seek additional information from human forecasters of operational forecasting agencies about the forecast and forecast quality beyond what is available publicly. For this objective, verification information should include comparative performance among different forecasts from different sources or based on different data sources, models and forecast systems. Such information should preferably be stratified for the environmental conditions that are being forecast (for example, past performance for flooding due to heavy rainfall on already wet soil from tropical storms and hurricanes). For the above, the verification task should address such user questions as "What is the basis for your latest forecast?" and "How confident are you of your forecast?" For the forecasters, the verification task should help them explain to the users how they arrived at their forecasts and what the significant sources of uncertainty may be.

Example III: Assessment of new techniques. A researcher has developed a new technique for correcting biases in streamflow forecasts. Her verification objective is to assess to what degree the bias-corrected forecasts were improved over the raw forecasts for possible operational implementation. She may decide on a set of metrics to measure accuracy. For both forecasts, she will design and carry out a hindcast (that is, reforecast) experiment for a period of multiple years. To determine if the bias-corrected forecast is improved over the raw forecast, she will compute skill scores of the bias-corrected forecast in reference to the raw (uncorrected) forecast. If the skill scores are significantly positive, she may conclude that the bias-corrected forecasts are better than the raw forecasts.

Example IV: Annual reporting. A government agency is tasked with routinely forecasting water levels for a stretch of a river used extensively for navigation. The error in the agency's river stage forecast may not exceed 10 and 20 cm at lead times of 12 and 24 hours, respectively. These performance indicators are assessed every calendar year. Each year, the agency prepares a report that shows the accuracy of the operational forecasts produced in the previous calendar year. The report also includes the same metrics for the preceding 10 years so that one may easily ascertain any long-term trends in forecast quality.

Example V: Ensuring that an ensemble forecast is probabilistically unbiased. A forecasting agency routinely produces ensemble streamflow forecasts. One of the users of the forecast would like to implement a formal decision-making process based on the probability of streamflow exceeding some threshold. The user wonders if he can justifiably assume that the probabilities derived from the ensembles are unbiased. He confirms that he is fine with all other forecast attributes of the ensemble forecast, and that he understands that the climatological ensemble forecast is perfectly reliable. The agency tasks a staff member to assess the reliability of the forecast. She collects the forecasts and the verifying observations from the operational archive. Using a verification tool, she produces rank histograms and reliability diagrams (see section 4.3.2). From these, she is able to verify if the forecast probabilities agree with observed relative frequencies and hence the ensemble forecast may be deemed probabilistically unbiased.

Example VI: Verification of run-time modifications by human forecasters. A forecast centre manually modifies forecast model inputs, parameters and states in an effort to produce the best possible forecast. The modifications are based on the forecasters' expert judgment as informed by routinely comparing model outputs with real-time observations. Every Monday, the previous week's forecasts are discussed in a plenary session. The discussion focuses on the degree to which forecasters' modifications have changed the quality of the forecasts.

These examples show that verification tasks may vary greatly in scope and depth, and that the verification process may look very different depending on the task at hand. The choice of the verification metrics and scores and the reference forecasts is likely to vary depending on the objective and the nature of the predictability of the predictand and the predictive skill in the forecast systems and processes. In addition, data sources may vary and, depending on the period of record available, hindcasting or reforecasting may be necessary.

5.2 Determining the audience

Verification information often has a distinct audience. An audience may be characterized in terms of its role, its level of understanding of the forecast process and its interest in forecast quality. One will have to make an estimate of the level of expertise that an audience has and adjust the verification products accordingly. Analogously to Werner et al. (2019), it is useful to distinguish several audiences of verification information: forecast users, forecasters, forecasting system developers and administrators.

5.2.1 Forecast users

Forecast users are those who are responsible for their own forecast-informed decisions. This audience may be interested in the degree to which forecasts improve their decisions. Forecast

users may vary from casual users (for example, members of the public who occasionally consult a forecast) to sophisticated users who use forecasts frequently and may have a formalized approach to using them in decision-making. The latter are likely to have a good understanding of the qualities they are looking for in a forecast, and hence the verification information will need to include measures of those qualities. Different users employ different decision support systems or processes of varying levels of complexity and sophistication. Some users may be able to utilize verification information explicitly in their decision-making process. In hydrological verification, insufficient location specificity often poses challenges for the users in utilizing forecast information. For example, the United States NWS issues official river forecasts only at the so-called "forecast points", which are usually collocated with stream gauge stations. If one lives away from such forecast points, the forecast information may or may not be relevant to one's location due to the very large control exerted by the local physiography on hydrological and hydraulic processes. The situation is similar with flash flood forecasts in that watches and warning are generally issued over large areas though flash flooding usually occurs only over a small fraction of the area. Similar challenges exist with verification information.

5.2.2 Forecasters

The role of forecasters is to provide the best possible forecasts. To do so they often adjust automatically produced forecasts. Such adjustments are based on the latest observations and expert knowledge of the hydrometeorology and hydrology of their service area as well as the forecasting system.

Verification information provides the forecasters with objective guidance on the quality and skill of the automatically produced forecasts in different forecasting situations. Verification serves a critical role in communicating the forecast information for user decision-making by aiding the forecaster to assess and convey the level of confidence or uncertainty in the forecast and to explain how the official forecast is arrived at. Verification information is also used to provide feedback to forecasters to improve the forecasts. Such feedback works best if the information is given as quickly as possible after the forecasts have been issued and the verifying observations have become available (see, for example, Azevedo and Bernard, 1995; Kulik and Kulik, 1988; White, 1968).

5.2.3 Modelers and developers

The role of forecasting system developers is to improve forecast quality by improving, newly developing or implementing various elements in the forecast system and in the forecast process. This may be carried out in a number of different ways, including selecting and configuring the most skillful NWP products, improving hydrological observations, and implementing or improving downscaling techniques, hydrological, hydraulic and water management models, DA or statistical processing techniques. Sometimes, the role of developer is taken on by members of the research community who are experts in these and other elements of a forecasting system and are intimately familiar with its configuration. Depending on whether an intermediate or end product is verified, different verification metrics may be necessary. Therefore, forecasting system developers tend to be well versed in the available metrics and their characteristics. It is hence common for forecasting system developers to also be the developers of forecast verification systems.

5.2.4 Administrators

The role of forecasting system administrators is to maintain, operate and improve the forecast systems, products and services. As part of that role, they are likely to want to know how well the forecasting system performs – both in terms of system operation (are forecasts produced, and in time?) and forecast quality. They are also likely to want to know how forecast quality improves over time. This information may be shared with a wider audience such as clients to whom the forecasts are supplied.

Administrators may not have a very intimate understanding of the technical nature of the forecasting process and may be interested only in the quality of the end product, and not that of the intermediate products such as precipitation forecasts or hydrological model simulations. The overall quality of the forecasting system may be expressed with a small set of key performance indicators agreed upon by the administrators and the stakeholders. Administrators are likely to understand the extent to which the key performance indicators are understood by third parties to whom an organization supplies its forecasts or from whom funding originates.

5.3 Selecting verification tools

Verification is generally a data intensive process for which software tools are used in most cases. One may use such tools to pair the forecasts with the verifying observations, compute summary metrics and skill scores, construct diagrams and visualize the results, among others.

Tools for verification include software packages that are specifically designed for forecast verification and generic tools that may be configured for forecast verification. Standalone software tools dedicated to verification include the Ensemble Verification System (Brown et al., 2010), the verif package (https://github.com/WFRT/verif) and the Model Evaluation Tools (https://dtcenter.org/community-code/model-evaluation-tools-met). Such tools come with ready-made verification functions that have been tested by many users and hence may be assumed to be free of functional errors (that is, the actual computation of a metric is correctly done). Such tools come at the expense of having to familiarize oneself with operating them and preparing the input data for ingestion by the tools. Generic tools include scripting languages such as R, Python and MATLAB and spreadsheet software such as LibreOffice Calc and Microsoft Excel. With these tools, verification data may be managed and used as inputs to functions written by the users. In-between solutions comprise libraries of verification functions that are written for use in generic tools such as the R verification library (NCAR – Research Applications Laboratory, 2015) and the Python verif library.

When deciding on a tool or tools for verification, one should consider the following questions:

- Is the tool able to perform the verification task?
- Are the information technology, computing capacity and data storage required for the tool (for example, software, scripting tools) available in or to my organization?
- Is the tool available for my computing platform or operating system?
- What is the level of maturity of the tool? Is documentation available? Is the tool supported? Does it have an active user community?
- Is the tool being actively developed? Is it open source so that, in principle, development and enhancement may continue by members of the community?
- What are the licensing conditions and associated costs, if any?

5.4 Collecting data

Once the verification task has been defined, it is necessary to identify the required data and to collect or produce them. The subject forecast may originate from an existing archive or may have to be retrospectively generated through hindcasting (that is, reforecasting). If forecast quality is to be compared against a baseline, the baseline forecast needs to be collected or produced as well. The forecasts may be verified against observations or against model simulations. Whereas the streamflow forecast reflects both the uncertainty in the input forcings and the hydrological uncertainty, the streamflow simulation reflects only the latter, thereby allowing decomposition of the two uncertainties (see Case 1 in Chapter 7).

In hydrological verification, large-to-extreme events are almost always of particular importance. Because they occur less frequently, their verification necessarily requires longer periods of record. In addition, the sample size requirements for verification of probabilistic forecasts are significantly larger than those for single-valued forecasts. For location-specific verification of ensemble streamflow forecasts of large-to-extreme events, multidecadal periods of record are usually necessary (Brown et al., 2014a, 2014b).

Forecasts to be verified often reside in an archive of operationally produced forecasts which reflect various changes that have been made over the period of the archive to the NWP and hydrological models. For example, the NWP models may have been improved over time, leading to improvements in the forecast forcings for the hydrological models. Similarly, model physics, calibration, DA or postprocessing components of the hydrological forecasting system may have been improved as well. Hence, verification of a multidecadal archive of operationally produced forecasts is not likely to reflect the skill of the forecasts generated from the current forecast systems and processes. To obtain large enough samples of forecasts from the current or a recent version of the forecast system, retrospective generation of forecasts, or hindcasting, is likely to be necessary. For this reason, hindcasting is often critical to large-sample hydrological verification, particularly in verification of probabilistic forecasts of large events.

Large-sample hydrological hindcasting necessarily requires a multidecadal reforecast dataset of the input variables (typically precipitation and temperature) from a fixed, or frozen, version of the NWP model. Such a dataset, however, may not exist or be readily available. Even if a reforecast dataset of the input variables is available, large-sample ensemble streamflow hindcasting for many locations is a resource-intensive undertaking. Therefore, it is particularly important that approximate sample size requirements be assessed for verification of rare events as part of the planning and design of the verification task. It is very possible that the above estimates may necessitate adjustments to the task.

5.5 Preparing data

For the verification tool to ingest the collected or generated data, they may have to be cast into a specific format first. For example, the verif tool requires data to be stored as a text file in a bespoke format or as a NetCDF-CF file. Creation of such files may require the use of another tool. In such a case, one will need to thoroughly assess the quality of the reformatted data to ensure their integrity with respect to the original data.

Often, probabilistic or ensemble forecasts are collapsed into ensemble mean or median forecasts and verified as single-valued forecasts by themselves or comparatively with other single-valued forecasts. If the verification tool does not support such conversion, the users will have to make the conversion themselves as part of the data preparation process.

Once forecasts and observations are collected, they must be paired (that is, the forecasts must be associated with their verifying observations). Depending on the verification software used, one may have to perform the pairing oneself or the tool may perform this task. For example, the verif tool requires paired data as input, whereas the EVS accepts separate files for forecast and observation data in multiple formats and pairs them automatically based on the user-chosen options.

Conceptually, pairing is a simple process, but in practice one may encounter various issues related to time stamping, time zones, time steps, forecast generation and valid times, and time scales of the forecast and observation (for example, instantaneous, time-averaged or accumulated). Sometimes, it may not be possible to achieve the initially desired pairing due, for example, to mismatches in forecast and observation valid times. In such cases, only approximate pairing may be possible. Once pairing has been completed, one will need to thoroughly assess the quality of the paired data to ensure their integrity with respect to the original forecast–observation pairs.

Most verification tasks require the full dataset to be subsampled for stratified verification or conditional verification. Predictability of precipitation and, to a lesser extent, temperature, varies significantly by season (for example, warm versus cool, wet versus dry), which leads to seasonally varying predictability of streamflow. Hence, it is common to stratify the forecast–observation pairs for different seasons and carry out verification for each season. In other cases, one may want to condition verification on some event of interest such as the observed peak streamflow exceeding 200 m^3/s or the observed 2 m air temperature being below freezing. Such conditioning requires the data to be subsampled prior to inputting to a verification tool or, if the tool allows for it, within the verification tool. Some verification tools, such as the EVS, will stratify paired data according to forecast lead time automatically. If the tool does not have such functionality, users will need to perform stratification themselves.

Verification of streamflow forecasts should preferably be carried out for individual locations since local physiography exerts great control on hydrological and hydraulic processes at the catchment scale. If it is necessary to increase sample size for verification of less frequent events, one may have to sort forecast points into hydroclimatologically similar groups that are more likely to share similar predictability and predictive skill. Such "pooling" of forecasts for increased sample size must be carried out with care to avoid mixing disparate forecasts. For example, pooling forecasts for headwater and downstream locations or flow volume forecasts for the snowmelt and dry seasons together is likely to produce distorted and misleading verification scores (Hamill and Juras, 2006).

Catchment size is an important factor for streamflow forecast quality, because larger catchments tend to be more predictable due to spatio-temporal smoothing of the hydrological processes. Hence, grouping forecasts from catchments of very different sizes is likely to produce confounding verification results. For large-scale events such as hurricanes and tropical storms, pooling may lead to biased verification results for large-to-extreme events. In such cases, the forecast–observation pairs for adjacent basins are likely to be correlated, and hence the effective sample size from pooling is likely to be smaller than the nominal sample size. When pooling data from hydroclimatologically similar catchments, one may first assess similarity among different basins by lowering the discharge threshold and comparing the verification results from pooling smaller observed events with those from location-specific verification. If the results differ significantly, the pooled locations are not likely to share similar predictability and predictive skill for larger events.

5.6 Computing verification statistics

Once the forecasts and observations have been paired and subsampled as necessary, verification metrics may be computed. Summary metrics are usually expressed as single numbers. Along with the values for the metrics, it is necessary to make available the metadata which describe the forecast used, the observation used, the period of record, and any subsampling or resampling used. The verification task may require skill scores to be computed, in which case the metadata should also identify the reference forecast used.

Often, the verification task requires uncertainty bounds to be computed. A commonly used technique for this is bootstrapping (Efron, 1979), which computes the metric many times, each time randomly subsampling the forecast–observation pairs with replacement. Bootstrapping yields an empirical distribution of the metric from which uncertainty bounds may be estimated (see Example 3 in Appendix A for a hands-on example using the EVS). If uncertainty bounds are not estimated, it is a good practice to provide the sample size information as a proxy. If the focus is on verification of large-to-extreme events, it is a good practice to identify the forecast–observation pairs for notable events by their names (for example, Hurricane Agnes in 1972 or Hurricane Harvey in 2017 in the United States). Many users of forecast information use such events as references.

Several verification measures are expressed as diagrams with standardized layouts. As described in Chapter 4, such diagrams include the ROC curve, Talagrand diagram (or rank

histogram), reliability diagram, attributes diagram and discrimination diagram. Most verification software tools will directly save these diagrams as image files, and in some cases, such tools will make available the values used to compose them. As with the metrics with numerical results, such diagrams need to be accompanied by metadata, some of which are often included in the diagram itself.

5.7 Key points

- Several steps are generally necessary to produce useful verification results. Each step involves several choices and decisions which, if not well thought out in advance, may lead to avoidable trials and errors.

- The high-level steps include: defining the verification objectives, determining the audience, selecting verification tools, collecting data, preparing data (including pairing), and computing verification statistics.

- The verification process may vary greatly depending on the verification task. It is hence very important to clearly define the verification objectives before starting.

- The verification audience determines how verification information may be presented. Possible audiences include forecast users, decision support staff, forecasters, forecasting system developers and administrators.

- Verification is generally data intensive and hence software tools are almost always used. Care should be taken to select a tool or tools that will meet the verification objectives and fit the resources available to the organization and to those who will use them. In addition, the tools should be supported, maintained and updated well into the future.

- The data required for verification may be readily available or may have to be produced through hindcasting (that is, reforecasting), which has potentially significant resource implications.

- The collected or generated data may have to be cast into a specific format for ingestion by the verification tool. Forecasts and observations must be paired for verification. Depending on the verification task, stratification, subsampling or conditioning of the full dataset may be necessary.

- Once the forecasts and observations are paired and subsampled as necessary, verification metrics may be computed. Several verification measures are expressed as diagrams with standardized layouts. Often, the verification task requires uncertainty bounds to be computed.

CHAPTER 6. VISUALIZATION OF VERIFICATION INFORMATION

In the verification, decision and response process, verification information will have to be communicated to a decision-making body. This chapter concerns the visualization of verification information that originates from the technical process of verification and is then communicated for the verification-informed decision (that is, the first arrow in Figure 23).

The verification results must be presented such that they address the questions posed in the verification task. Verification results may be laid out in graphical or tabular form. While most verification tools have the option to display verification results, the user of verification tools may want to use some other tools for more advanced display capabilities. For the above, the user will have to confirm that the output from one tool is ingestible by the other. In some cases, the verification questions may be partly answered by displaying the raw data themselves. The raw data are often shown graphically as time series or scatter plots. While many verification tasks may require summary metrics only, visualization of the underlying data is also often helpful in interpreting the summary metrics (as well as to perform quality control on the data).

6.1 Visualization of forecast verification

Verification information may include raw data, technical diagrams, summary metrics and metadata. For effective and clear communication of graphical verification metrics, graphing best practices should be followed to avoid the information being misinterpreted by the recipient. There is ample literature on this topic; a good starting point may be found in Tufte (2013).

6.1.1 Raw data

A first impression of the quality of the forecasts can be obtained from exploring the raw data for forecasts and their verifying observations. This includes the original forecast hydrographs, time series and scatter plots.

Deterministic, ensemble and probabilistic forecasts can all be displayed as hydrographs in some form. Usually, a hydrograph is depicted by a single forecast only (that is, a forecast valid from the initialization time to the end of the forecast horizon or the maximum lead time). For ensemble and probabilistic forecasts, often only the mean or the median (or some other quantile) forecast is displayed. It is important to note that, strictly speaking, such plots are not hydrographs in that they do not represent plausible realizations of a hydrograph, whereas the individual ensemble members do. For this reason, care should be taken when visually comparing ensemble mean or quantile forecast “hydrographs” with the verifying observed hydrograph, as the former are likely to appear smoother than the latter. Figure 24 shows an example display of an ensemble forecast hydrograph and the verifying observed hydrograph.

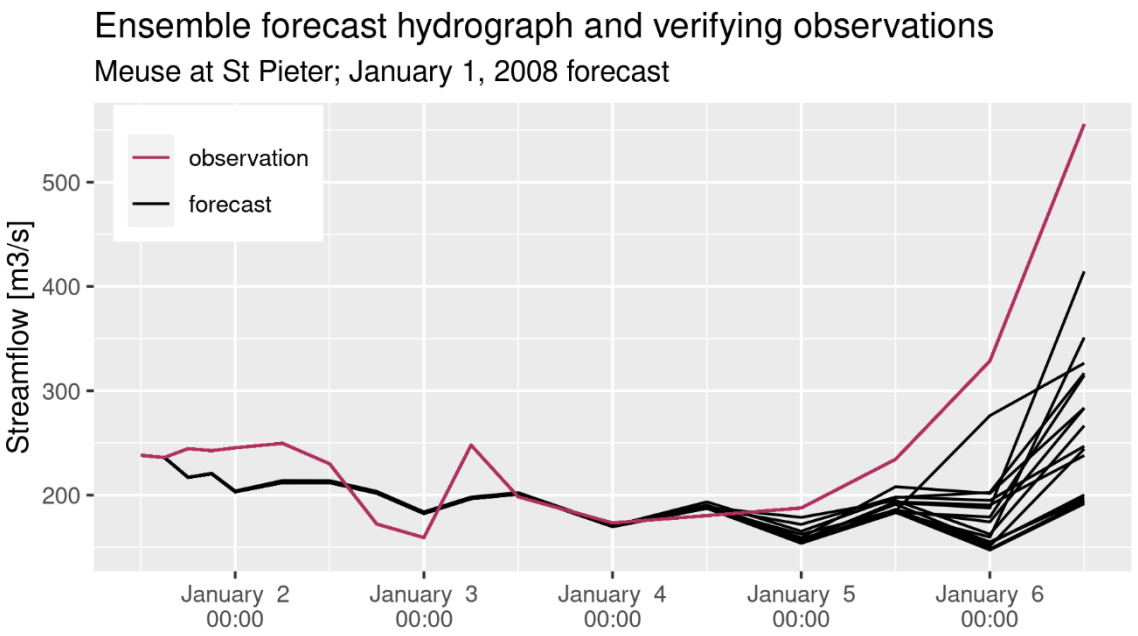


Figure 24. An ensemble forecast with verifying observations plotted as hydrographs

Time series plots may be used to visualize multiple forecasts with their verifying observations. Often, these plots show forecasts for a single lead time only (see Figure 25 as an example).

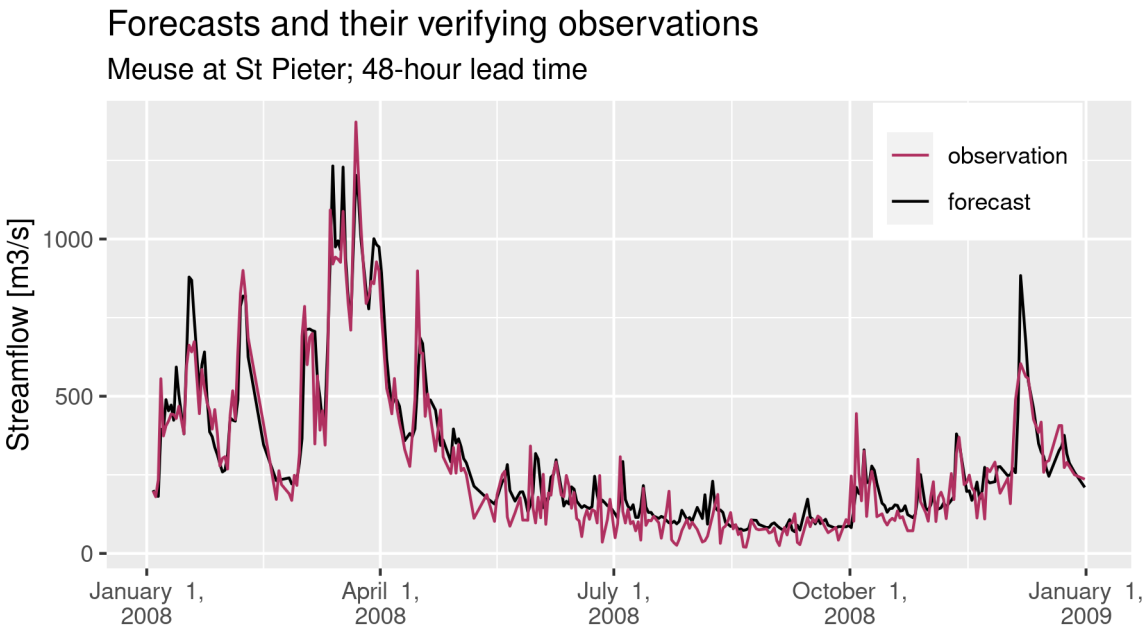


Figure 25. A timeseries plot of forecasts made 48 hours ahead and their verifying observations for a single location. Note that this plot shows a composite of multiple forecasts, hence one could have opted to not show this as a continuous time series.

Scatter plots may be used to visualize multiple forecasts and their verifying observations. Each of these forecast–observation pairs is represented by a single point in the scatter diagram (see Figure 26 for an example), or by multiple points for ensemble forecasts. In the latter case, for each verifying observation there are multiple forecasts associated, each representing an ensemble member (see Figure 27 as an example).

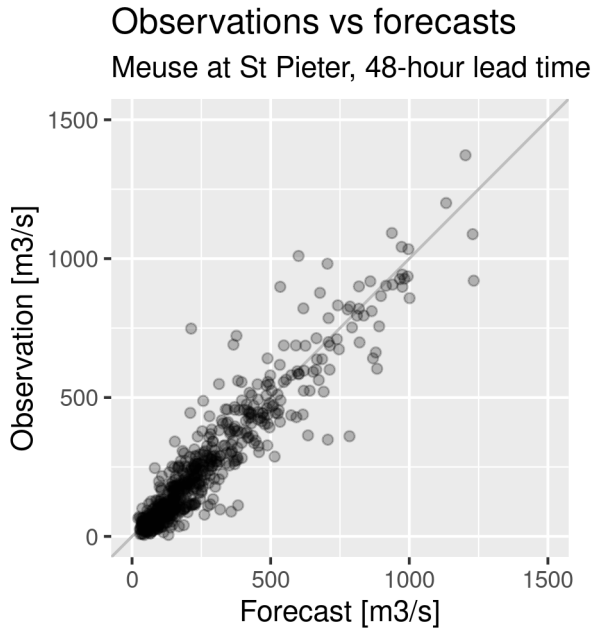


Figure 26. A scatter plot of forecasts made 48 hours ahead and their verifying observations for a single location

Sometimes, depending on the distribution of forecasts and observations, many points may crowd the same locations in a scatter plot. This prevents the reader from accurately assessing the distribution of the forecast and observations. This issue is often overcome by using transparent symbols that are shaded differently to indicate the number of forecast-observation pairs that each symbol represents. This is done in Figure 26 and Figure 27; light grey circles indicate single points whereas dark grey and black circles indicate multiple points in each circle.

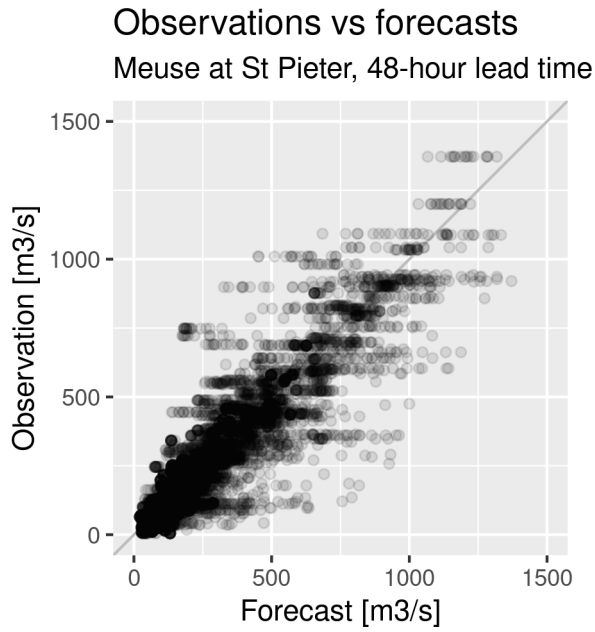


Figure 27. A scatter plot of ensemble forecasts and their verifying observations

Scatter plots may also be used to display forecasts at multiple lead times versus verifying observations. For such plots, one may colour-code the markers to differentiate different lead times (see Figure 28 as an example).

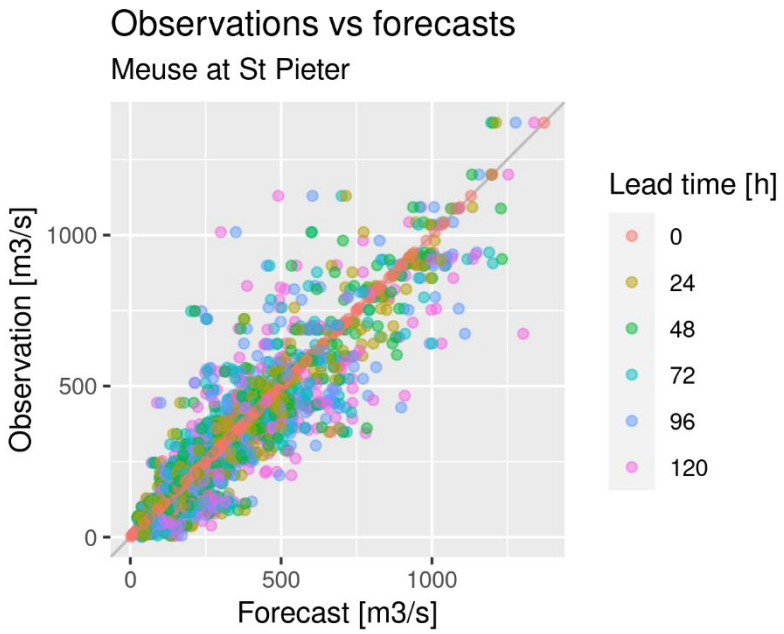


Figure 28. A scatter plot of forecasts at multiple lead times with their verifying observations

6.1.2 Technical diagrams

Some verification measures are displayed as diagrams with predefined layouts. They include the reliability diagram, attributes diagram, ROC curve, rank histogram and discrimination diagram; section 4.3.3 contains a description of these. An example is shown in Figure 29. Note that the plot also displays metadata, including the location, the forecast lead time and the event definition.

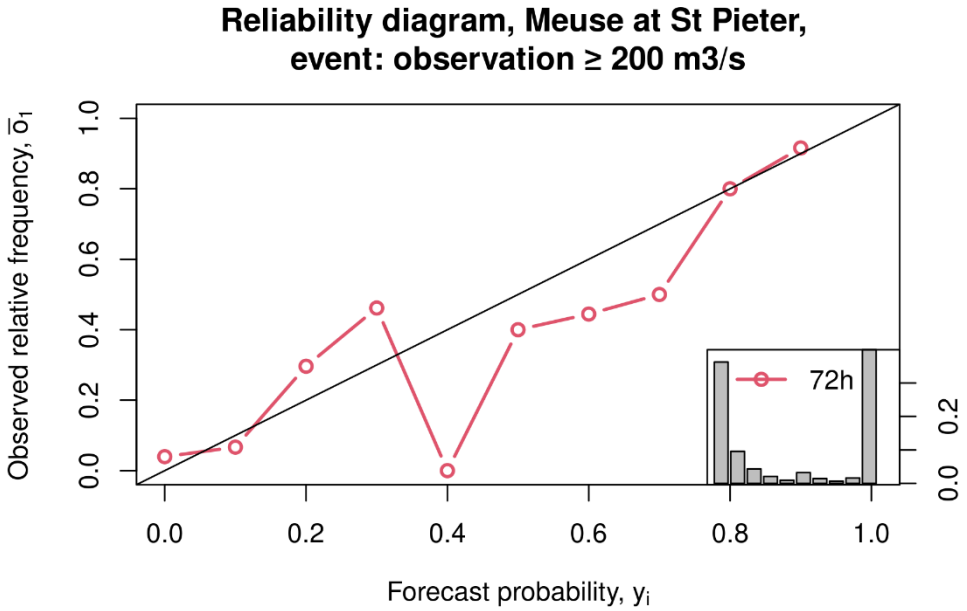


Figure 29. Reliability diagram for 72-hour forecasts for a single location. This plot was drawn using the R verification package (NCAR – Research Applications Laboratory, 2015).

6.1.3 Summary metrics

Summary metrics may be communicated in various ways, including in descriptive text, in tabular form, in a graph or on a map. Tables, graphs and maps allow for the summary metrics to be shown as a function of some property of the forecast. For example, many verification studies report summary metrics as a function of forecast lead time for a particular location. Figure 30 shows an example of this. Maps show summary values as a function of space (Figure 31 shows an example of this). Tables may do both, that is display summary metrics as a function of both location and lead time.

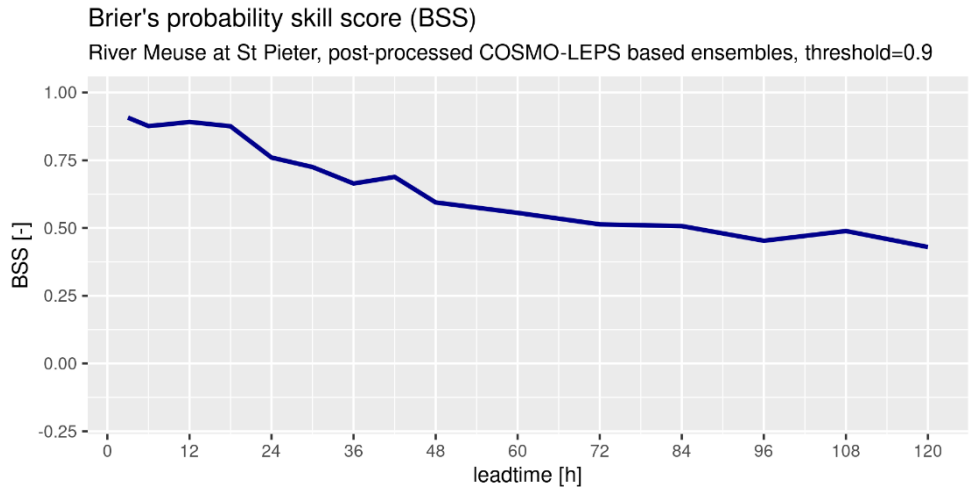


Figure 30. Brier skill score (BSS) as a function of lead time

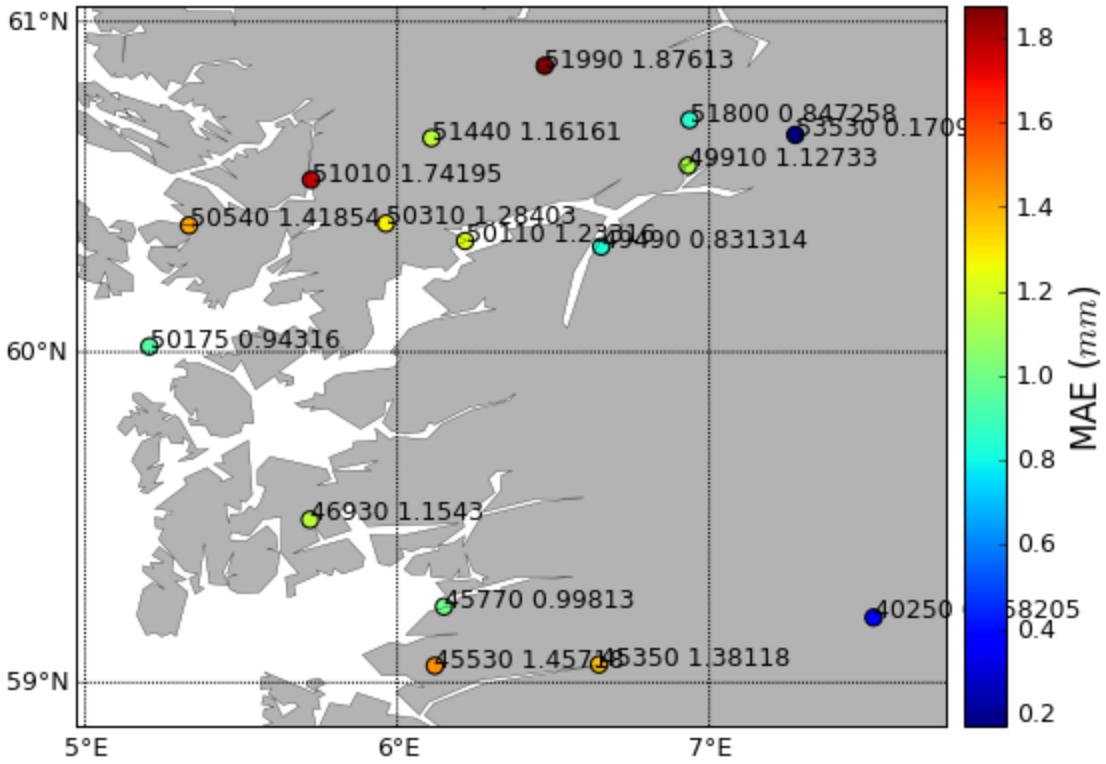


Figure 31. Mean absolute error of a precipitation forecast, presented as a map. The numbers alongside the colour-coded points show a location identifier and the value of the metric.

Source: Example taken from https://github.com/WFRT/verif/wiki/Plotting-options.

If forecasts are regularly verified over time, the verification data may be presented as a time series. This will add additional dimensions to the data, namely the time at which the verification was done and the period it pertained to. Figure 32 illustrates this. Here, the verification is carried out every few months. Over time, this yields a time series of verification data which allows one to assess the changes in (a particular type of) forecast quality over time.

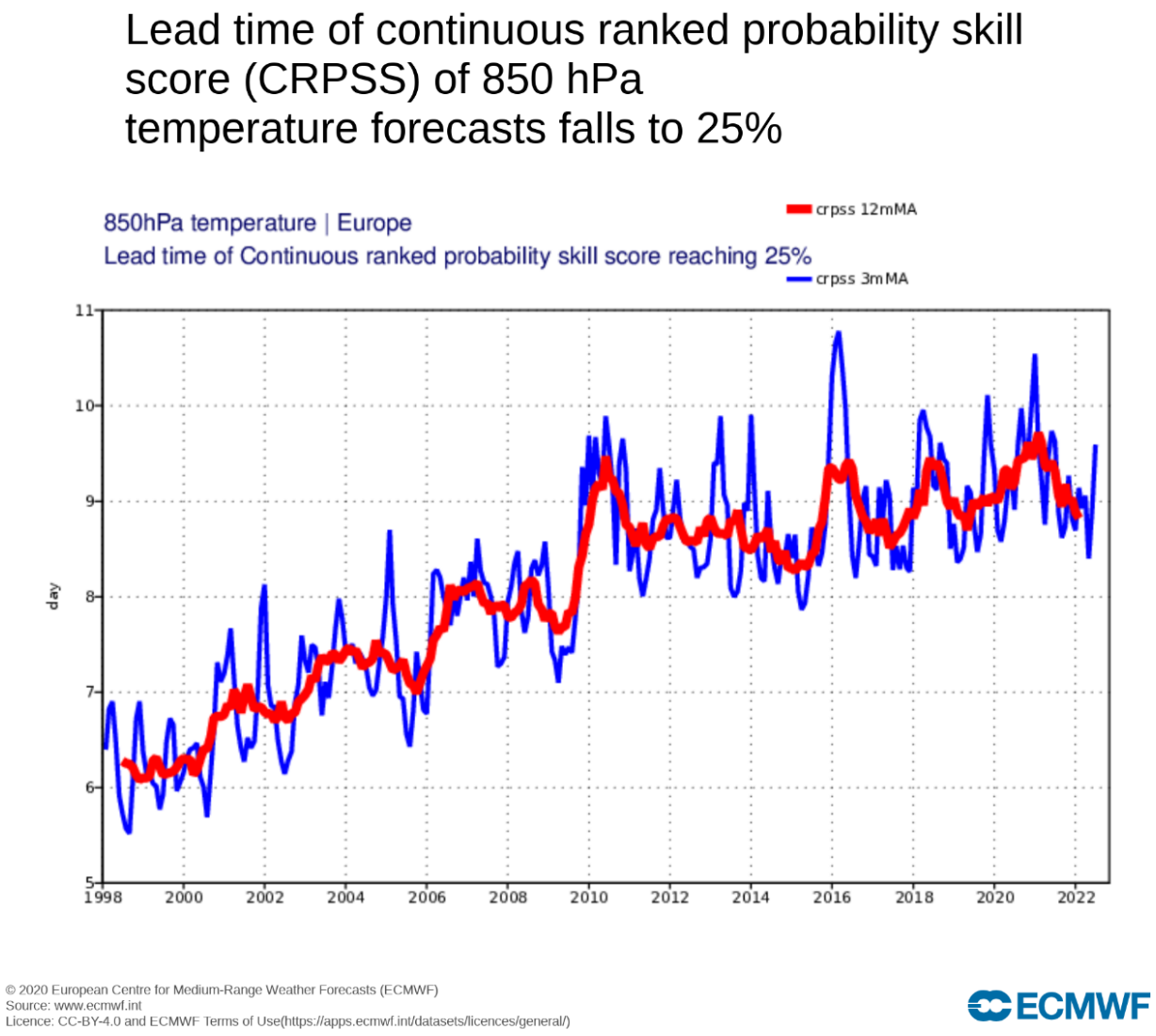


Figure 32. Lead time at which the CRPSS reaches the value of 0.25, as a function of time. The plot shows that over time (from 1998 through 2022), the CRPSS of the forecast has reached 0.25 at an increasingly longer lead time.

Source: https://charts.ecmwf.int/products/plwww_m_eps_tpcrpssreach_ts?area=Europe

It may not be assumed that all verification users will be familiar with how a particular metric is computed. This unfamiliarity limits the user in being able to properly interpret the value of the metric. If one suspects that users may not be familiar with how the metric is computed, it is a good practice to include a reference to the exact composition of the metric or, better yet, a guide as to how the metric should be interpreted. Figure 33 shows an example; below the graph showing the metric value (in this case, the CRPSS), a brief description of the metric is given and a link to a more detailed definition is provided.

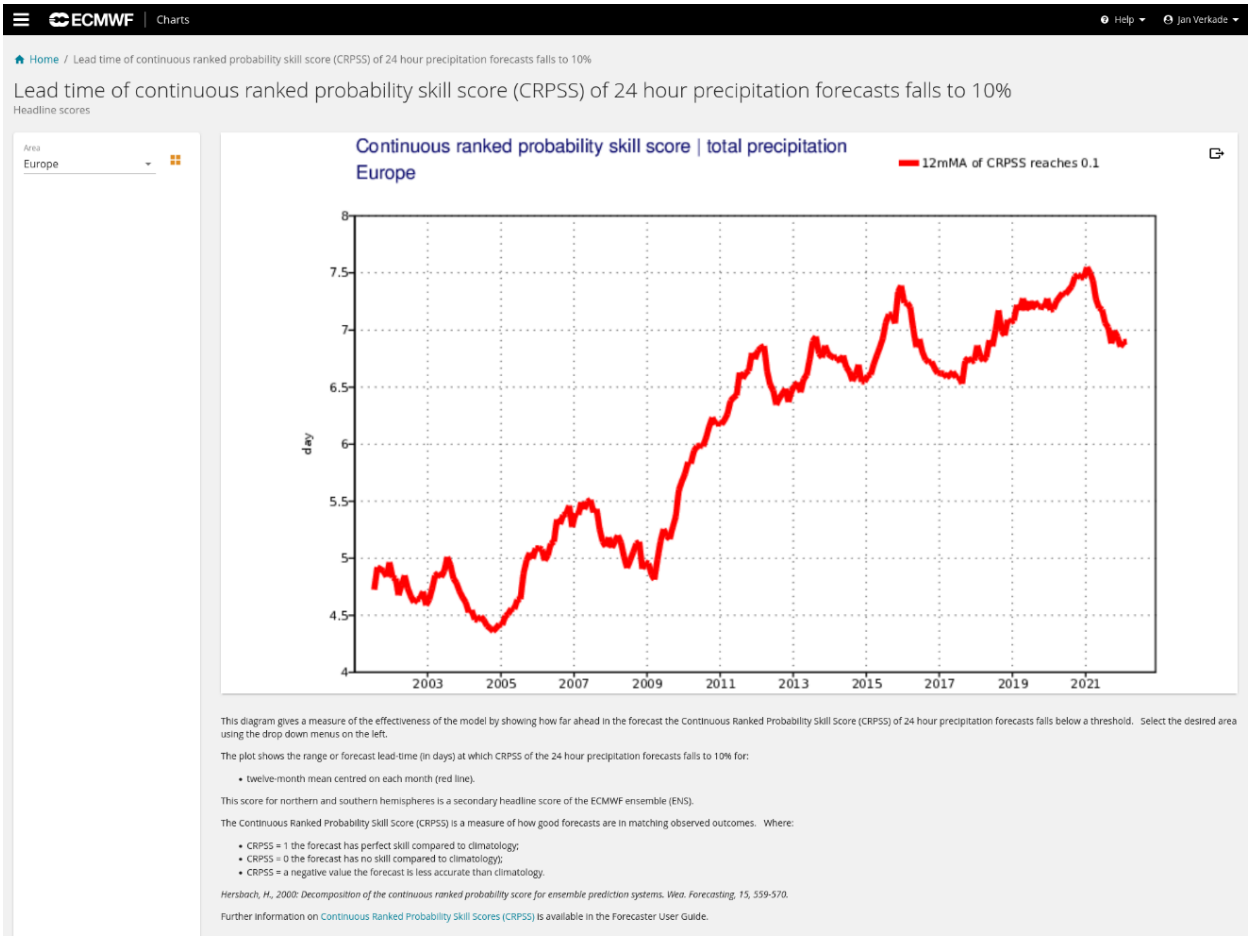


Figure 33. Web page that shows verification information, a brief description of the metric shown and a reference to a more elaborate definition of that metric

Source: Example taken from
https://charts.ecmwf.int/products/plwww_m_eps_tpcrpssreach_ts?area=Europe

6.2 Metadata

Verification information is characterized by a large amount of metadata. This metadata describes the specific set of forecasts that may be chosen according to the time period of verification interest (for example, forecasts produced in the calendar years 2012 and 2013). The forecasts may be conditioned on various attributes, including issue time or valid time. The verification measure may pertain to a specific subset of the data based on the exceedance or non-exceedance of a threshold value by either the observation or the forecast. The verification may pertain to a single location or to a set of locations. It may pertain to a single lead time or possibly be aggregated over multiple lead times. All of these are choices made in the verification process and thus become properties of the verification data. The number of these properties quickly adds up and, by construction, so does the number of combinations thereof.

Example: Upon verification of a set of hydrological forecasts for the Meuse River, summary metrics (Table 9) are available for four forecast locations, two forecast products and two different parameters (water level and streamflow rate). Verification was done separately for 17 different lead times up to 120 hours at 3-hour to 6-hour intervals. The verification was done separately for the entire dataset and for various subsets of the data. Here, the latter are denoted with a P value. For continuous forecasts, these P values indicate a subset defined by the observation exceeding, for example, the 90% quantiles of the empirical distribution. For those metrics that are used to verify a categorical forecast, the P value denotes the threshold. A total of seven summary metrics are computed, and a bootstrapping procedure yields a lower, a central and an upper estimate.

This yields summary metric values for each of the four locations, two forecast products, two parameters, seven lead times, two P-values, seven summary metrics and three estimates (a grand total of 11 424 values). Each of the rows in the table contains a summary metric value which can be communicated to the verification audience. No single (effective) visualization will show all, and hence a selection must be made, and this selection must be communicated to the audience.

Table 9. First few rows of a database table containing verification metadata and verification data. The actual value of a summary metric is stored in the "value" column. All other columns comprise metadata items, and for each of these, multiple values exist in the database. For example, the database comprises multiple locations, multiple products that are verified, and so on.

	location	product	parameter	lt	metric	P	estimate	value
1	H-MS-SINT	eps-dressed	Q-fas	3	bss_sco	0.9	central	0.90747
2	H-MS-SINT	eps-dressed	Q-fas	6	bss_sco	0.9	central	0.87636
3	H-MS-SINT	eps-dressed	Q-fas	9	bss_sco	0.9	central	0.88345
4	H-MS-SINT	eps-dressed	Q-fas	12	bss_sco	0.9	central	0.89144
5	H-MS-SINT	eps-dressed	Q-fas	18	bss_sco	0.9	central	0.87564
6	H-MS-SINT	eps-dressed	Q-fas	24	bss_sco	0.9	central	0.76009
7	H-MS-SINT	eps-dressed	Q-fas	30	bss_sco	0.9	central	0.72496
8	H-MS-SINT	eps-dressed	Q-fas	36	bss_sco	0.9	central	0.66434
9	H-MS-SINT	eps-dressed	Q-fas	42	bss_sco	0.9	central	0.68869
10	H-MS-SINT	eps-dressed	Q-fas	48	bss_sco	0.9	central	0.59448

Verification information may be communicated using various media, including paper or screen. Most figures and tables will be 2-dimensional. This allows only a subset of metadata properties to be shown, hence the verification information will have to be conditioned on some of the remaining properties. The visualization will have to clearly indicate on which properties the data are conditioned. A trivial example is "location"; a summary metric versus lead time graph may pertain to a single location (see, for example, Figure 30). This location will need to be communicated in the graph title, caption or similar metadata. In similar fashion, the graph's metadata will need to identify for which variable (for example, streamflow or water level) and for which forecast product it was made.

It may be helpful for a user to explore verification data using a tool that makes it possible to condition the verification results on metadata properties. This requires the verification results to be stored in a structured manner, that is, in a table like the one in Table 9 or in a database, and a tool that can access and visualize that data. Such a tool may be a scripting language such as MATLAB, R or Python, or a bespoke tool. Figure 34 shows an example of such a bespoke tool. This "dashboard" based on Shiny (Chang et al., 2023) accesses a large R dataframe which is essentially a database. The dashboard allows a user to compose his or her own visualizations by navigating the available metadata items.

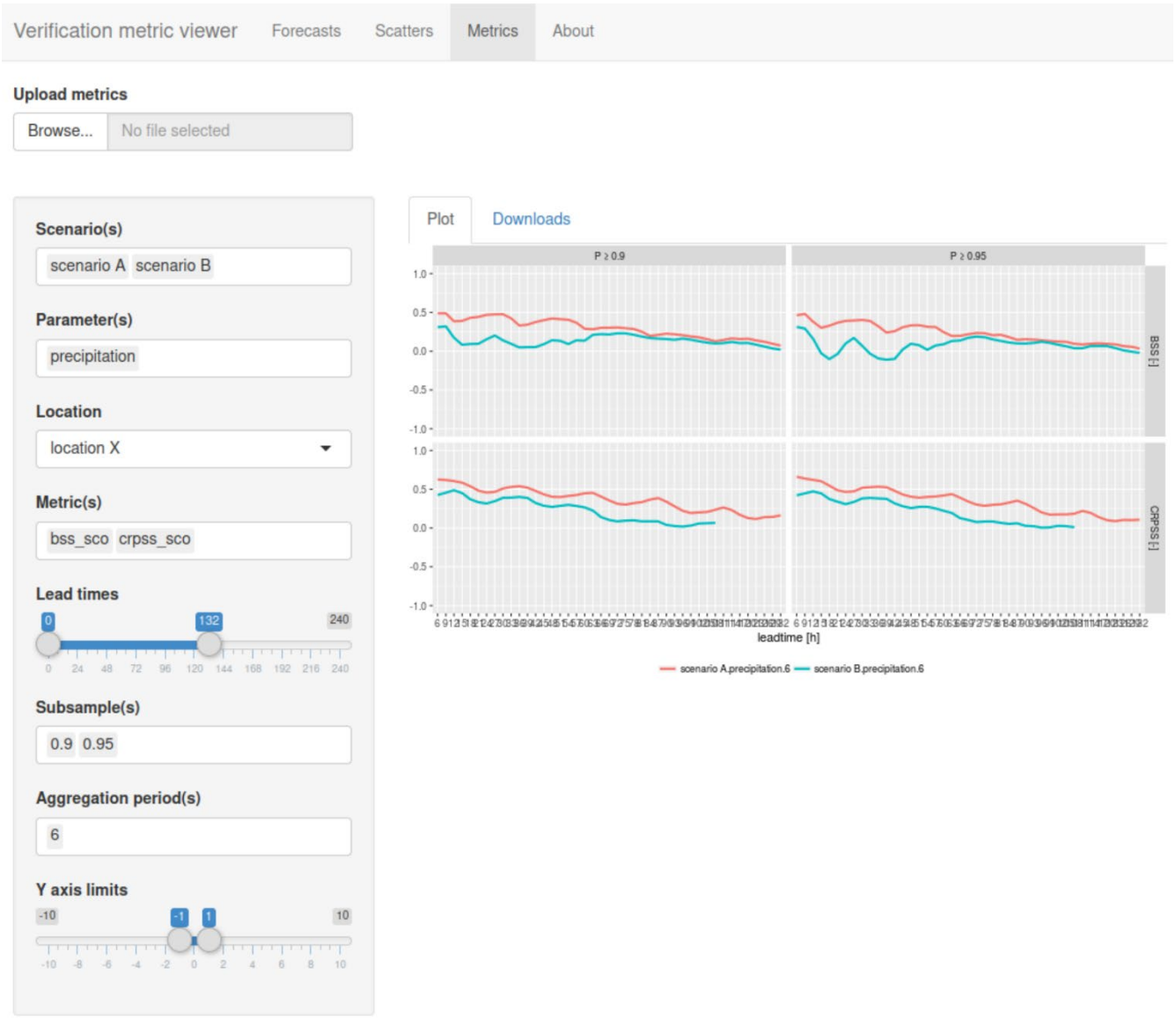


Figure 34. Example of a tool (unpublished) that allows a user to compose bespoke visualizations of summary metrics

6.3 Key points

- Verification information may include raw data, technical diagrams, summary metrics and metadata.
- Verification information generally includes a large amount of metadata.
- Verification can yield a large number of verification statistics. These cannot all be shown in a single plot or table, and hence accompanying metadata will have to identify the "dimensions" that are not explicitly communicated in the plot or table.

CHAPTER 7. CASE STUDIES

This chapter presents several cases of streamflow forecast verification, as listed in Table 1. Though presented individually, the cases collectively encompass a wide range of analysis one may encounter in the verification of hydrological forecasts. The reader may consider mixing and matching the various elements described herein based on the verification task at hand and introduce additional elements as necessary. Appendix A provides hands-on examples of streamflow forecast verification, as listed in Table 2.

Verification statistics are a condensation of the comparison between the forecast and the verifying observation wherein the full information content resides. It is therefore a good practice to visualize the raw data in hydrologically and hydrometeorologically interpretable terms, such as in scatter plots, box-and-whisker plots (for ensemble forecasts), hydrographs and hyetographs, so that one may tangibly relate the verification statistics with the physical world. Such a practice is in fact often necessary to perform quality control on the data. If the verification statistics and the raw data do not seem to agree, it may be necessary to examine the data more closely and vet the verification process before continuing.

7.1 Case 1. Verification of single-valued streamflow forecast with uncertainty decomposition

Example 1 in Appendix A provides a hands-on exercise for this case using the EVS. Case 1 verifies single-valued streamflow forecasts for the New River at Galax (GAXV2), Virginia, United States, which is the outlet of a 2 397 km^2 headwater basin in the service area of the Ohio River Forecast Center (OHRFC) of the United States NWS. Throughout this chapter and Appendix A, five-character identifiers such as GAXV2 indicate the NWS forecast points. The streamflow forecast was generated with the Sacramento soil moisture accounting (SAC) (Burnash et al., 1973) and UHG routing (Chow et al., 1988) models. The SAC was forced by the Global Ensemble Forecast System (GEFS) v12 (Guan et al., 2019) ensemble mean precipitation reforecast. The period of record is 1989 to 2004.

To decompose the total uncertainty into input and hydrological uncertainties (see Chapter 2), streamflow simulation was also generated with the SAC using observed mean areal precipitation (MAP) as input. Hydrological uncertainty is further decomposed into the IC uncertainty and the rest (that is, the sum of parametric, structural and other uncertainties) (see Chapter 2) by updating the ICs of the hydrological models in a real-time mode. The automatic state updating procedure used in this case is based on the assimilation of observations of streamflow, MAP and mean areal evapotranspiration via adaptive conditional bias-penalized ensemble Kalman filter (Seo et al., 2022; Shen et al., 2022b).

The focus of this case is on verification of streamflow forecasts and uncertainty decomposition for significant events. Therefore, only significant events were identified from the observed streamflow data and the resulting partial-duration time series was used. A significant event is defined as a hydrograph whose peak flow exceeds 175 m^3/s. This corresponds to approximately the 90th percentile of observed flow at this location (for reference, the flood flow is 937 m^3/s). Once a significant event was identified, the beginning and ending hours of the event were determined by subtracting and adding 5 days from and to the first and last hours of the partial-duration series exceeding the threshold, respectively. The choice of 5 days was made to include most of the significant portions of the primary baseflow response. The primary verification measure used is the RMSE, which reflects multiple forecast attributes (see Equation 48) and is representative of the overall forecast quality. Figure 35 shows the RMSE versus lead time and the uncertainty decomposition results.

In Figure 35, the solid, dashed and dotted lines represent the RMSE of the streamflow forecast forced by the GEFS ensemble mean precipitation forecast, streamflow simulation forced by the observed MAP, and streamflow prediction forced by the observed MAP but with state updating,

respectively. The recurrence of peak RMSE values every 24 hours is due to the diurnal variations and forecast cycles in the skill of precipitation forecasts (Brown et al., 2014a, 2014b; Siddique et al., 2015). The difference between the solid and dashed lines represents the input uncertainty. The difference between the dashed and dotted lines represents the initial condition (IC) uncertainty as inferred via state updating (that is, adjusting the ICs in real time to bring the simulated flow in line with the observed in the immediate past) (Shen et al., 2022b). Figure 35 indicates that the bulk of the input uncertainty rises quickly within the first day or so of lead time, hydrological uncertainty is very large when only significant events are considered, and large IC uncertainty exists over short lead times.

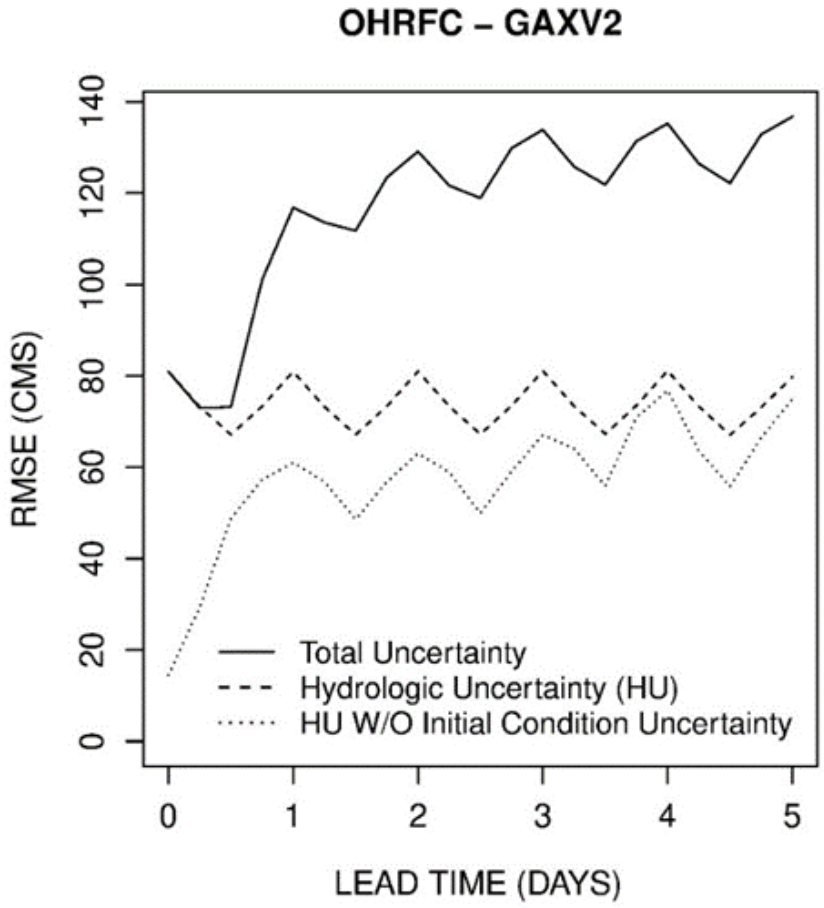


Figure 35. RMSE versus lead time of streamflow forecast for GAXV2 and its decomposition into input and hydrological uncertainties. CMS: cubic metres per second.

One may decompose the MSE into contributions from bias in the mean, bias in variability and deficit in correlation (Murphy and Winkler, 1987; Nelson et al., 2010):

$$MSE = (\mu_X - \mu_Y)^2 + (\sigma_X - \sigma_Y)^2 + 2\sigma_X\sigma_Y(1 - \rho_{X,Y}) \quad (48)$$

where μ_X and μ_Y denote the mean of the forecast and verifying observation, respectively, σ_X and σ_Y denote the standard deviation of the forecast and verifying observation, respectively, and $\rho_{X,Y}$ denotes correlation between the forecast and verifying observation. The above decomposition is not related to the CR or LBR decompositions described in Chapter 3. Figure 36 shows the MSE decomposition result for the total uncertainty where *Comp1*, *Comp2* and *Comp3* represent the first, second and third components in Equation 48, respectively. Figure 36 shows that all three components contribute significantly to total uncertainty for this location and that, overall, the third component is the largest contributor. Improving correlation, $\rho_{X,Y}$ in Equation 48, across all forecast horizons generally requires improving NWP, hydrological modelling and model calibration.

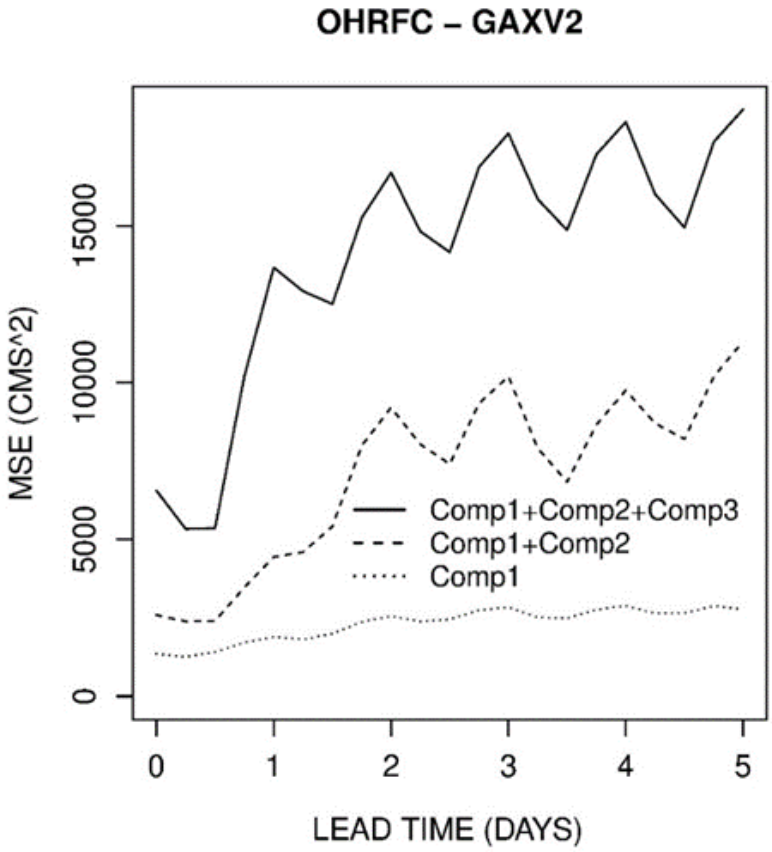


Figure 36. MSE decomposition for streamflow forecast for GAXV2. CMS: cubic metres per second.

7.2 Case 2. Comparative verification of multiple single-valued streamflow forecasts

Case 2 is an adaptation of Jozaghi et al. (2021) and verifies single-valued streamflow forecasts from multiple forecast systems. The forecasts included are the singled-valued River Forecast Center (RFC) forecast, ensemble mean forecast generated with the Hydrologic Ensemble Forecast Service (HEFS) (Demargne et al., 2014), single-valued medium-range forecast from the National Water Model (NWM) (Graziano et al., 2017), and ensemble mean forecasts forced by precipitation and temperature forecasts from the GEFS (Toth and Kalnay, 1997; Cui et al., 2012) and the North American Ensemble Forecast System (NAEFS) (Zhu and Toth, 2008). The NAEFS forcing forecast combines the ensemble forecasts from the Meteorological Service of Canada and the United States NWS (Zhu and Toth, 2008). Except for the NWM forecast, all other streamflow forecasts are generated at the NWS Middle Atlantic RFC (MARFC). The forecasts are verified for the Delaware River Basin (DRB), which has a drainage area of about 33 000 km^2 and includes parts of the states of Delaware, New Jersey, New York and Pennsylvania in the United States. For details of the models and the forcings used, refer to Jozaghi et al. (2021).

Due to the very short period of record of January 2017 to October 2020, it was not possible to verify for individual forecast points. Instead, all forecast points within the basin were pooled together for aggregated verification. Because the magnitude of discharge and the forecast error characteristics vary from one location to another, basin-aggregated verification does not represent the individual forecast points equally. To address this issue, one might consider some form of normalization based on catchment size and possibly other hydroclimatological attributes. Such an approach, however, does not allow for quantitative interpretation of verification measures expressed in physical units, such as RMSE, and was not pursued in this case study.

In general, streamflow at headwater locations is significantly less predictable than that at downstream locations due to the large input and hydrological uncertainties associated with rainfall-runoff and hillslope routing processes. Predictability is larger for downstream locations owing to the spatio-temporal aggregation and smoothing over large areas and significantly smaller hydrological uncertainty associated with channel routing. To account for this variation in predictability with different hydrological controls, the forecast points were grouped into 11 headwater and 15 downstream locations and verified separately. One could also consider stratification with respect to season to account for the variation in predictability with hydroclimatological controls. Such additional stratification was not used in this case study due to insufficient sample size.

Flood forecasting is arguably the most important service provided by the MARFC, and hence verification of large flows is of particular interest. Accordingly, the forecasts are verified conditionally on the verifying observed flow exceeding the 95th percentile at the respective forecast points, in addition to being verifyied unconditionally by including all pairs of forecasts and verifying observations within the DRB. In general, flood flows are much larger than the 95th percentile flows. Larger thresholds, however, would render the sample size too small to yield useful results. The choice of the 95th percentile is hence a compromise between keeping the sampling uncertainty to a tolerable level and still being able to assess comparative predictive skill for high flows. The verification measures used are the MSE, ME and CORR (see section 4.3). Figures 37 through 40 show the results.

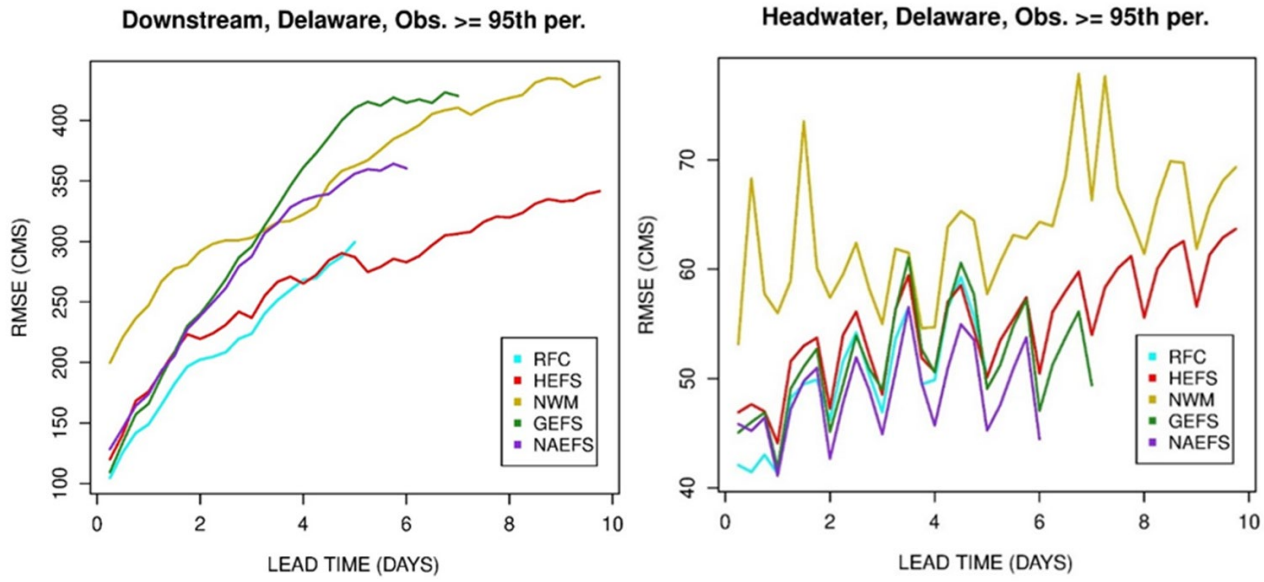


Figure 37. RMSE versus lead time of different streamflow forecasts conditional on the verifying observation exceeding the 95th percentile flow for downstream (top) and headwater (bottom) forecast points in the DRB

In the graphs in Figure 37, the maximum lead time varies from one streamflow forecast to another due to the different forecast horizons in the forcing forecasts (see Jozaghi et al., 2021 for details). Due to the daily forecast cycle at MARFC, not all streamflow forecasts are generated immediately after the forcing forecasts become available. This operational constraint translates, in effect, to penalties in lead time of 6 hours for the GEFS and NWM forecasts and 12 hours for the NAEFS forecast.

Figure 37 shows the conditional RMSE versus lead time of single-valued and ensemble mean streamflow forecasts for all downstream (top) and headwater (bottom) forecast points in the DRB. The sample size ranges from 453 to 517 and from 751 to 857 for headwater and downstream locations across all lead times, respectively. As expected, the downstream points are much more predictable, with the conditional RMSE plateauing at about 4 days into the future. The headwater results show peak RMSE values recurring every 24 hours due to a

combination of the diurnal cycle in the skill of precipitation forecast and some forecasts verifying rather poorly for a very small number of very large observed flows. Because forecasts are issued daily, multiple forecasts are verified against the same very large observed flow across the forecast horizon, resulting in the cyclical pattern. Figure 37 indicates that, overall, the RFC single-valued and NAEFS ensemble mean forecasts have the smallest RMSEs for high flows for downstream and headwater locations, respectively.

Figure 38 shows the RMSE versus lead time for all ranges of verifying observed flow, or the unconditional RMSE, for headwater locations. The unconditional RMSE results for the downstream locations are qualitatively similar to the conditional results shown in the top panel of Figure 37. The sample size for the unconditional results ranges from 8 750 to 9 750 and from 14 150 to 16 000 for headwater and downstream locations across all lead times, respectively. With all ranges of flow, the skilful lead time is increased to about 3 and 5 days or longer for headwater and downstream locations, respectively. Figure 38 shows that the relative unconditional performance among the forecasts for headwater locations is quite different from the relative conditional performance seen in the bottom panel of Figure 37, and that the single-valued RFC and ensemble mean HEFS forecasts have the smallest unconditional RMSE.

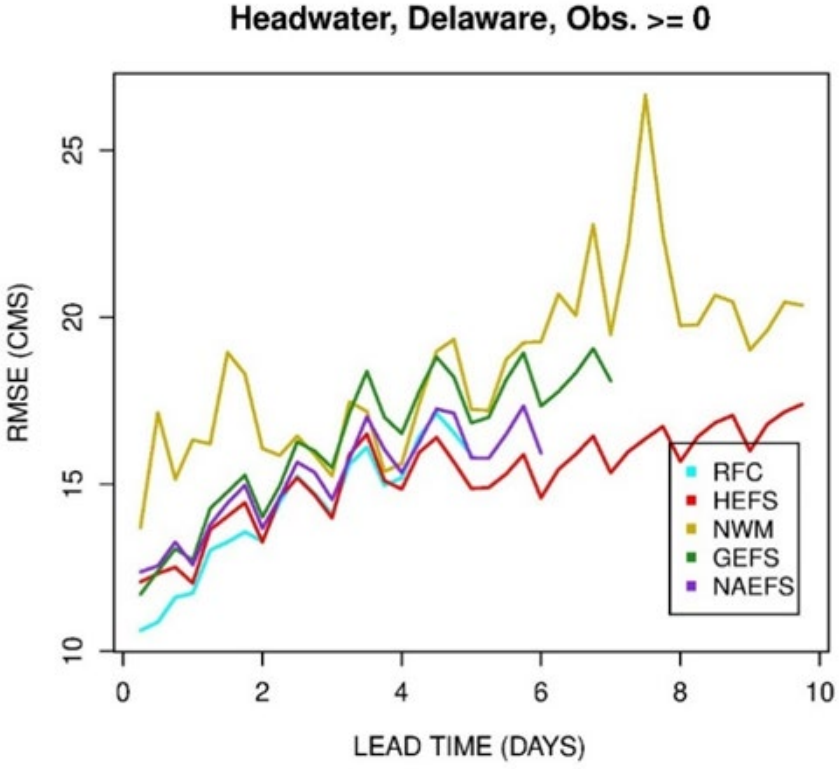


Figure 38. Same as bottom panel of Figure 37 but for all ranges of flows

Figure 39 shows the conditional (top) and unconditional (bottom) ME for downstream locations. Those for headwater locations are qualitatively similar. The bottom panel of Figure 39 shows that the HEFS forecast is unconditionally unbiased, whereas the GEFS and NAEFS forecasts are unconditionally biased high (that is, over-forecast in the mean sense), with the bias increasing with lead time. Figure 39 illustrates a common challenge in streamflow forecasting, where smaller bias in the GEFS and NAEFS forecasts for high flows is achieved at the expense of large high bias in flows of all magnitudes, whereas the unconditional unbiasedness of the HEFS forecast is achieved at the expense of large low bias in high flows. Jozaghi et al. (2021) describe a multimodel streamflow prediction approach to address the above via composite conditional bias-penalized linear regression which explicitly considers type II conditional bias.

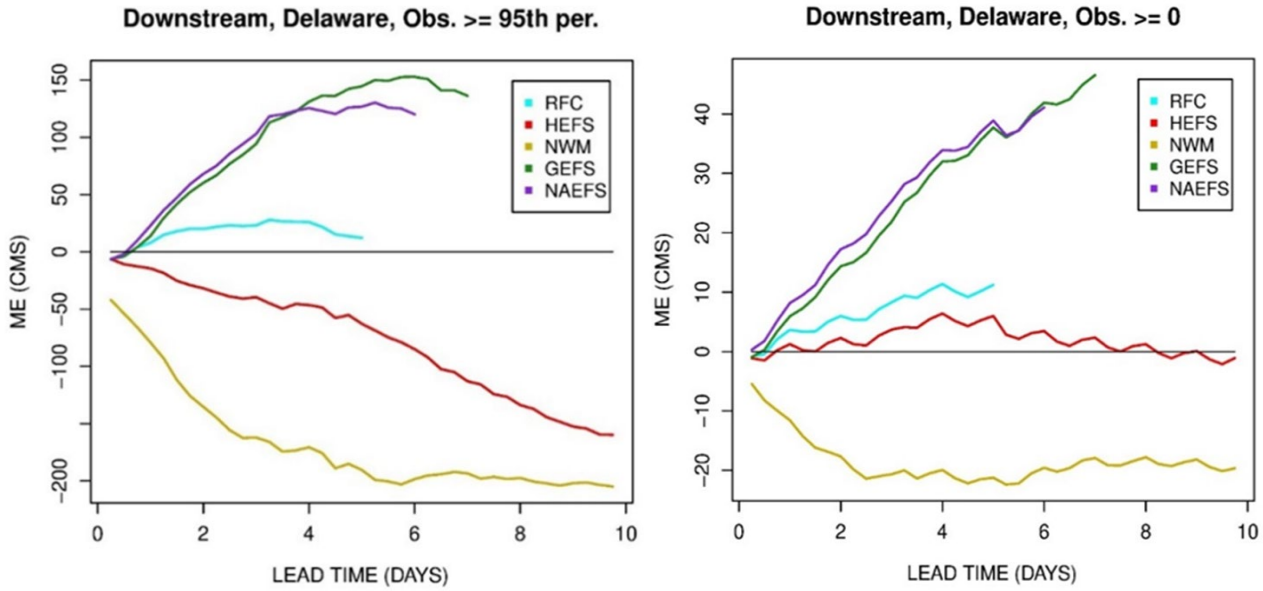


Figure 39. ME versus lead time of different streamflow forecasts for downstream forecast points in the DRB conditional on the verifying observation exceeding the 95th percentile flow (top) and for all flows (bottom)

Figure 40 shows the conditional CORR for downstream (top) and headwater (bottom) locations. The unconditional CORR results are qualitatively similar. Figure 40 shows that, at headwater locations, predictability is much smaller and the predictive skill diminishes much more quickly with lead time than at downstream locations. The NWM and HEFS forecasts show noticeably lower CORR for downstream and headwater locations, respectively, for attribution of which close examination of the raw data is necessary.

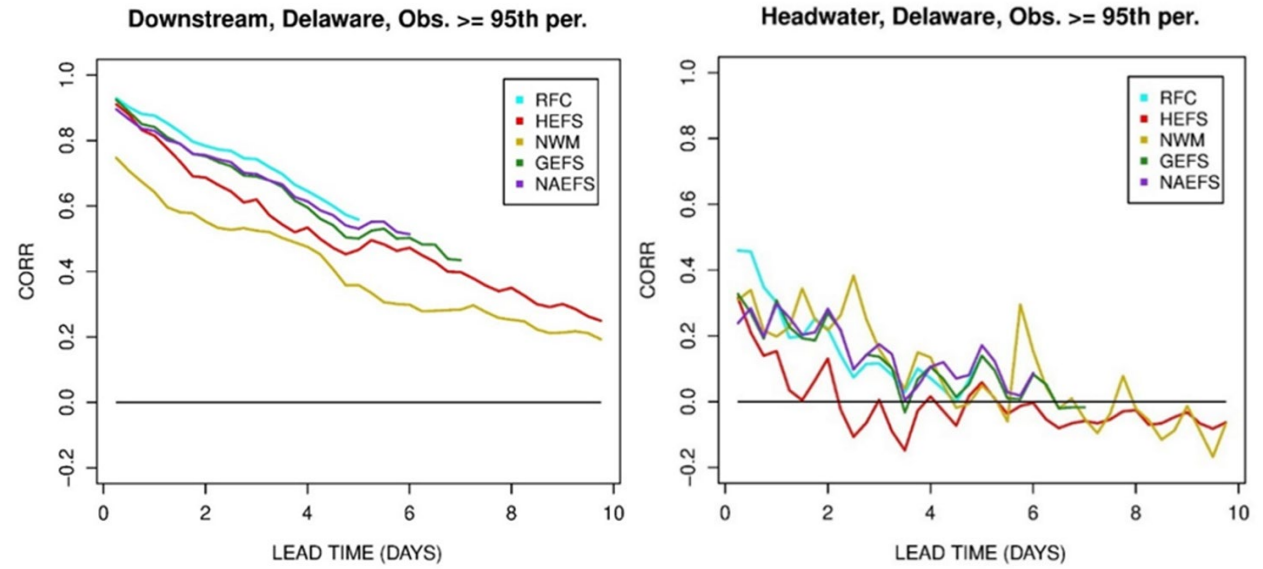


Figure 40. Same as Figure 37 but for CORR

7.3 Case 3. Verification of ensemble streamflow forecast for headwater and downstream locations

Example 2 in Appendix A provides a hands-on exercise with a smaller version of this case using the EVS. Case 3 assesses reliability, resolution, discrimination and sharpness (see Chapter 3 and Chapter 4) of the NAEFS ensemble streamflow forecast produced by the MARFC using the mean CRPS and its CR decomposition (Hersbach, 2000), reliability diagram (Hsu and

Murphy, 1986) with forecast frequency histogram, and ROC curves (Stephenson and Jolliffe, 2003) (See Chapter 4). Recall that NAEFS-forced ensemble mean streamflow forecasts were used in Case 2 for comparative verification of single-valued forecasts. The period of record is less than three years, which is too short for location-specific probabilistic verification of flood or near-flood flows. To increase the sample size, all 11 headwater and 15 downstream forecast points in the DRB are pooled together for forecast group-aggregated verification as in Case 2.

The mean CRPS is a representative measure of ensemble forecast quality, reflecting multiple attributes of ensemble forecasts (see Equation 49). As with the RMSE for single-valued forecasts, the mean CRPS is expressed in the same physical unit as the predictand itself. Figure 41 shows the mean CRPS of NAEFS-forced streamflow forecasts for downstream (top) and headwater (bottom) locations as generated with the EVS. In each graph, the "All data" curve represents the unconditional result based on all available forecast–observation pairs. The other four curves are conditioned on the verifying observation exceeding the 50th, 75th, 95th and 99th percentiles of observed flow. The sample sizes for the 95th and 99th percentiles are 2 301 and 456, respectively, for downstream locations, and 197 and 44, respectively, for headwater locations. The discharge values in m^3/s associated with the conditioning percentiles are shown in the legend. Flood flows are generally significantly larger than the 99th percentile. The use of higher percentiles, however, is often not possible due to insufficient sample size.

In Figure 41, the saw-tooth pattern in the 95th and 99th percentiles for the headwater locations is due to several forecasts verifying rather poorly for one or more very large observed flows as explained in Case 2. Because the observed flow associated with such a pattern is usually the maximum flow observed within the period of record, the amplitude of the pattern often provides a very useful indicator of sampling uncertainty at the highest thresholds. Figure 41 shows that, for the 95th and 99th percentiles, the mean CRPS for the downstream locations increases slowly with lead time, an indication of large predictability, whereas that for the headwater locations jumps up almost immediately at very short lead times, an indication of very limited predictability. Also note that, even though the 99th percentile flow for headwater locations is less than a fifth of that for downstream locations, the magnitude of the mean CRPS for headwater locations is comparable to that for downstream locations at very short lead times, in reflection of the very limited predictability at headwater locations.

Figures 42 and 43 show, for downstream (top) and headwater (bottom) locations, the RES and REL components of the mean CRPS, respectively. They are based on the CR decomposition of the mean CRPS into REL, RES and UNC, or into REL and potential CRPS (Hersbach, 2000) (see also section 4.3):

$$\mathrm{CRPS} = \mathrm{REL} - \mathrm{RES} + \mathrm{UNC} = \mathrm{REL} + \mathrm{CRPS}_{\mathrm{POT}} \qquad (49)$$

A smaller REL indicates more reliable ensembles (which is desirable), and a larger absolute value of RES means better resolution (which is also desirable). The RES component ($= \mathrm{UNC} - \mathrm{CRPS}_{\mathrm{POT}}$) is positive if the ensemble forecast is better than the climatological ensemble forecast (Hersbach, 2000). The UNC component reflects climatological uncertainties in the observations and does not relate to forecast attributes. The $\mathrm{CRPS}_{\mathrm{POT}}$ represents the mean CRPS achievable by calibrating forecast probabilities to be perfectly reliable (Hersbach, 2000). As with the CRPS, the smaller the $\mathrm{CRPS}_{\mathrm{POT}}$ is, the better the forecast is.

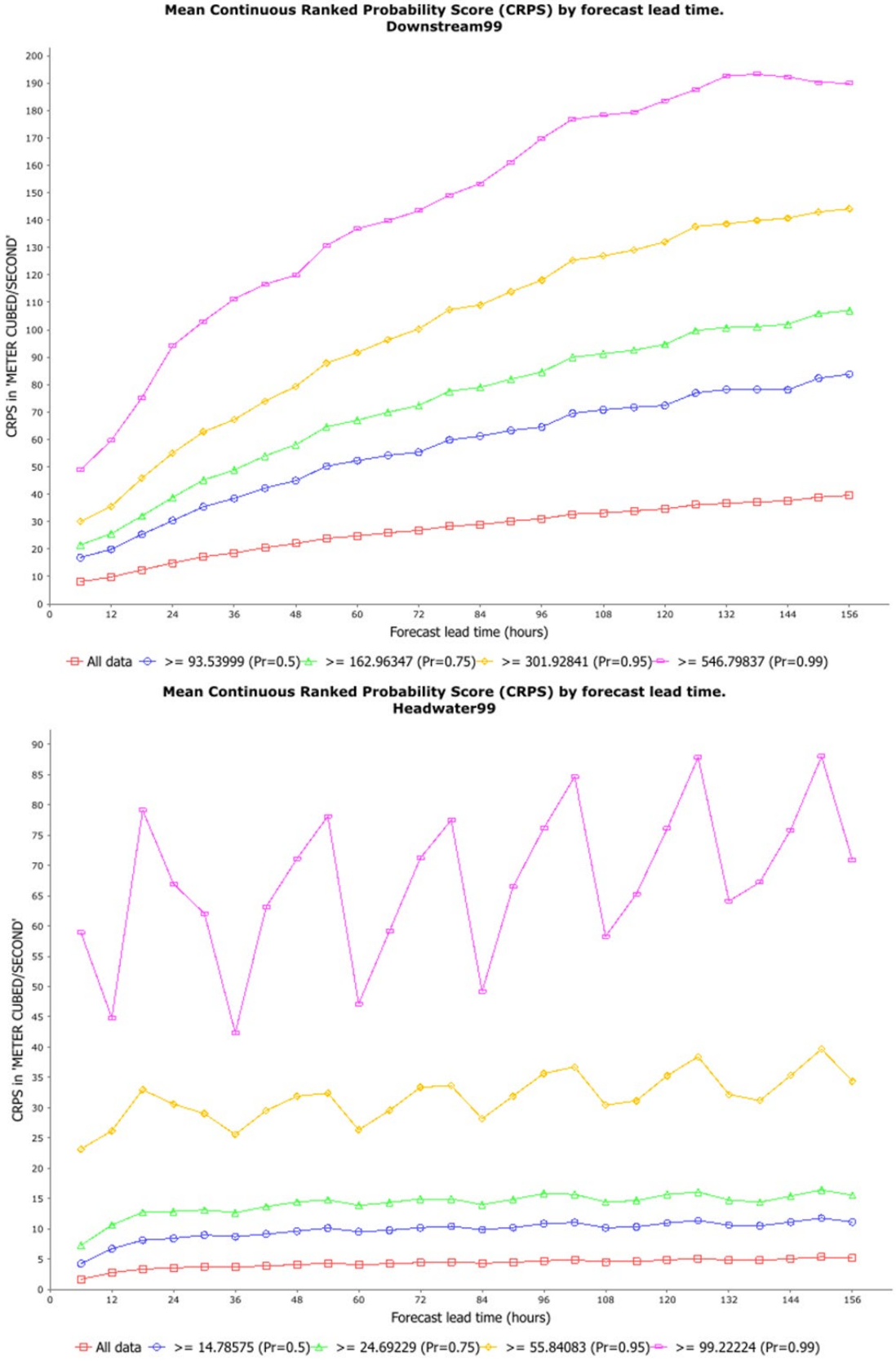


Figure 41. Mean CRPS of NAEFS-forced ensemble streamflow forecasts versus lead time for downstream (top) and headwater (bottom) forecast points in the DRB

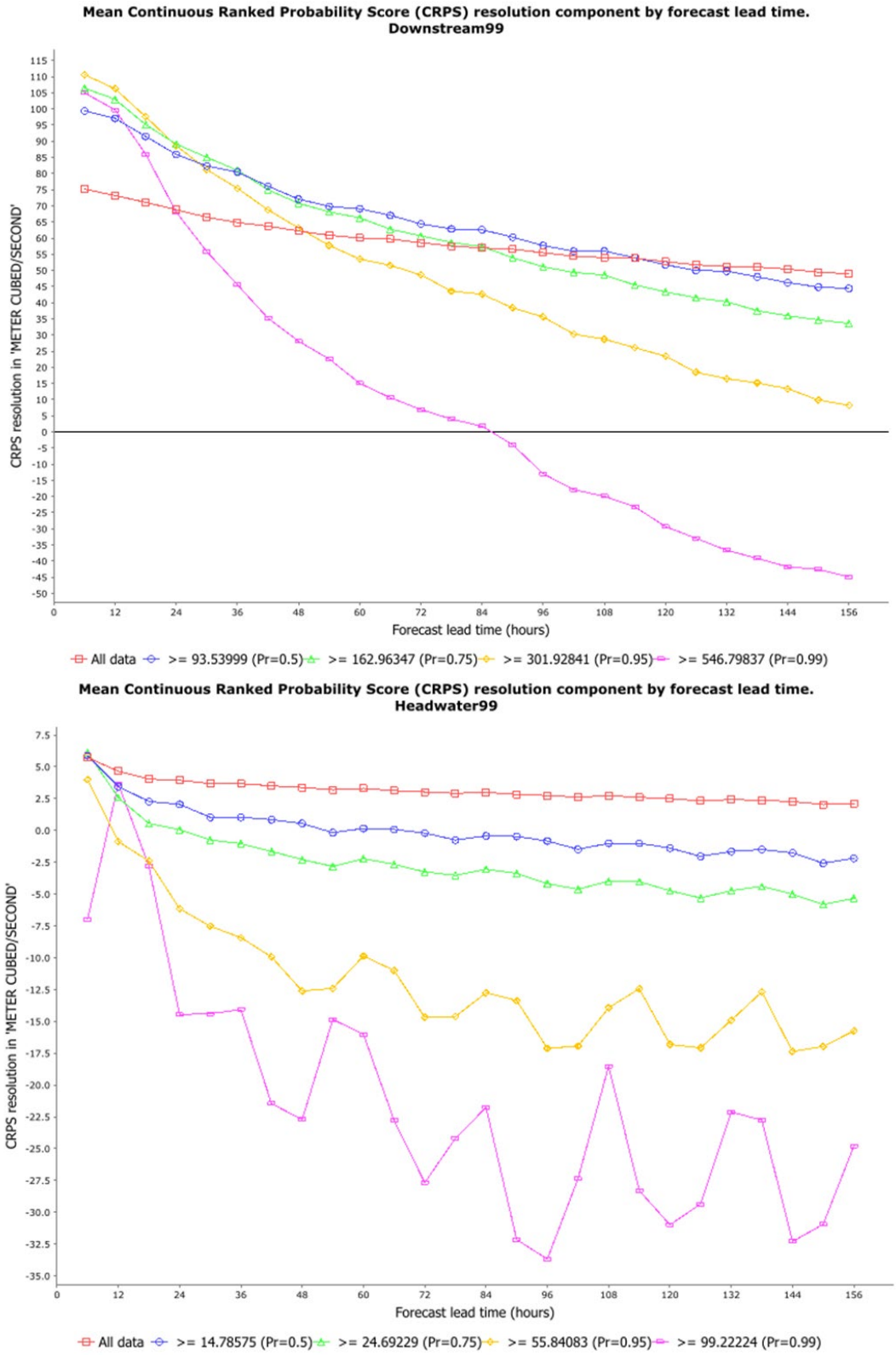


Figure 42. Same as Figure 41 but for the resolution component of the mean CRPS

Figure 42 shows that the NAEFS-forced ensemble streamflow forecasts have positive RES out to multiple days into the future for downstream locations for the 95th and 99th percentiles. However, for headwater locations, RES quickly turns negative, thus negatively contributing to the mean CRPS.

Figure 43 shows that the NAEFS-forced streamflow ensemble forecasts for downstream locations have similar levels of REL when conditioned on the 95th and 99th percentile observed

flows, but that they are significantly less reliable for headwater locations when conditioned on the 99th percentile flow versus the 95th percentile. The particularly unfavourable REL in Figure 43 at short lead times for headwater locations is due largely to significant low bias in the NAEFS-forced ensemble streamflow forecasts. The above observations indicate that the large mean CRPS for headwater locations is due to both unfavourable RES and unfavourable REL, the latter of which may potentially be improved by statistical post-processing with a much larger sample size.

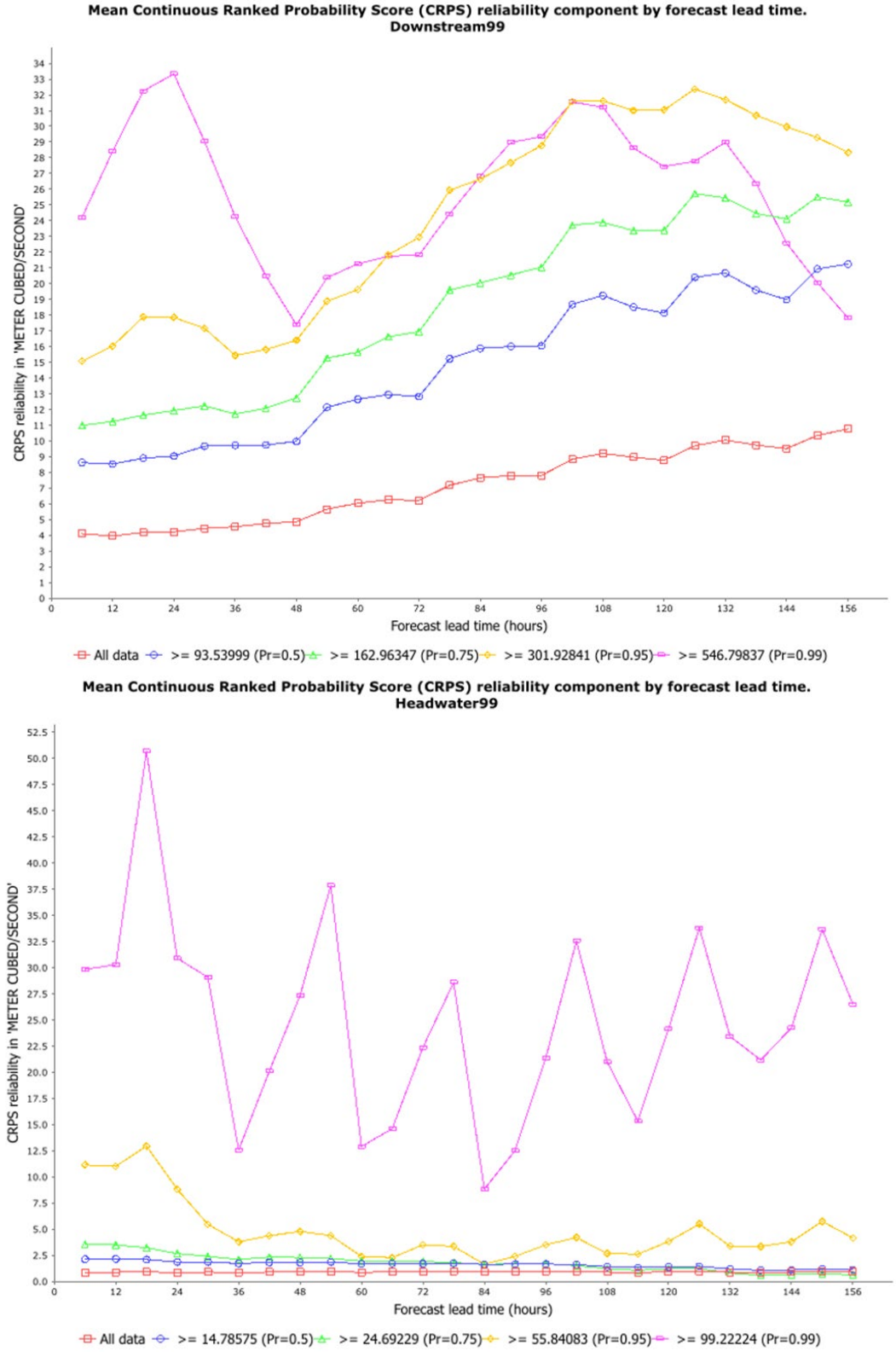


Figure 43. Same as Figure 41 but for the reliability component of the mean CRPS

The REL component in Equation 49 only assesses how closely in the mean sense the ensemble members mimic the cumulative probability distribution of the observed flows over all ranges, similarly to the rank histogram but accounting for ensemble width (Hersbach, 2000). Hence REL is a weaker test of reliability than the reliability component of the BS across a range of thresholds. To assess reliability in a stronger sense, one may use the reliability diagram as shown in Figure 44 for the NAEFS-forced ensemble streamflow forecasts at lead times of 18 (top) and 156 (bottom) hours for downstream locations. The figure shows that the ensemble forecast is generally reliable at very short lead times for downstream locations, but that reliability is reduced at longer lead times due to high bias (see Figure 39).

The bottom part in each panel of Figure 44 shows frequency histograms of the forecast exceedance probabilities associated with the respective thresholds. If the ensemble forecast is perfectly sharp (see Chapter 3 and Chapter 4), the histogram would show probability masses at 0 and 1 only. If the ensemble forecast is not very sharp, the histogram would appear shrunk towards the middle from all three sides (left, right and bottom). It is seen in the histograms that the ensemble forecast loses sharpness as the lead time increases. Compared to the longer lead time (bottom panel), the sample size for forecast probability of zero is larger and the forecast probabilities of 0.3 to 0.7 are smaller at the shorter lead time (top panel).

Figure 45 shows the ROC curves for NAEFS-forced ensemble streamflow forecasts at a lead time of 156 hours for downstream (top) and headwater (bottom) locations, respectively. The ROC measures the ability of the forecast to discriminate between events and non-events (see Chapter 3 and Chapter 4). Analogously to CORR being immune to linear bias in single-valued forecasts, ROC is insensitive to reliability of ensemble forecasts (that is, bias in probability space). The ROC explicitly reflects the forecast's ability or inability to detect events given the threshold of interest, such as flood stage. The ROC is hence a particularly useful measure for large-to-extreme events for which Type II (that is, false negative) error is important. In Figure 45, an event is defined for simplicity as the observed flow exceeding the threshold at any given time. However, any other event definition of arbitrary complexity can be used as long as it is possible to condition the forecast–observation pairs consistently.

Figure 45 shows that NAEFS-forced ensemble streamflow forecasts have significantly better discriminatory skill for downstream locations than they do for headwater locations, and that the skill is more sensitive to the choice of the threshold for headwater locations than for downstream locations. For headwater locations, the POD is only about 0.49 even when only 10% of the ensemble members exceed the 99th percentile flow, with a POFD of 0.05 (that is, the uppermost marker on the yellow line). For downstream locations, the matching POD is much larger at 0.95, with a POFD that is only marginally larger at 0.11, resulting in a much larger AUC than for headwater locations.

For single-valued forecasts (or, equivalently, perfectly sharp probabilistic forecasts), the mean CRPS and ROC reduce to the MAE and a single point of (POD, POFD), respectively. They hence allow for comparative assessment of skill between ensemble and single-valued forecasts. Caution must be exercised in such a comparison, however, as probabilistic verification necessarily favours ensemble forecasts; because they are perfectly sharp ensemble forecasts, single-valued forecasts are penalized for "sticking their neck out" whereas ensemble forecasts are rewarded for hedging their bets. On the other hand, if the mean CRPS of the ensemble forecast turns out to be larger than the MAE of a comparable single-valued forecast or if the (POD, POFD) of the single-valued forecast is found to lie outside (that is, above or to the left) of the ROC curve, it indicates that the ensemble forecast system may be deficient in certain areas compared to the single-valued forecast system (for example, uncertainty modelling, forcings, model physics, initialization, calibration or postprocessing).

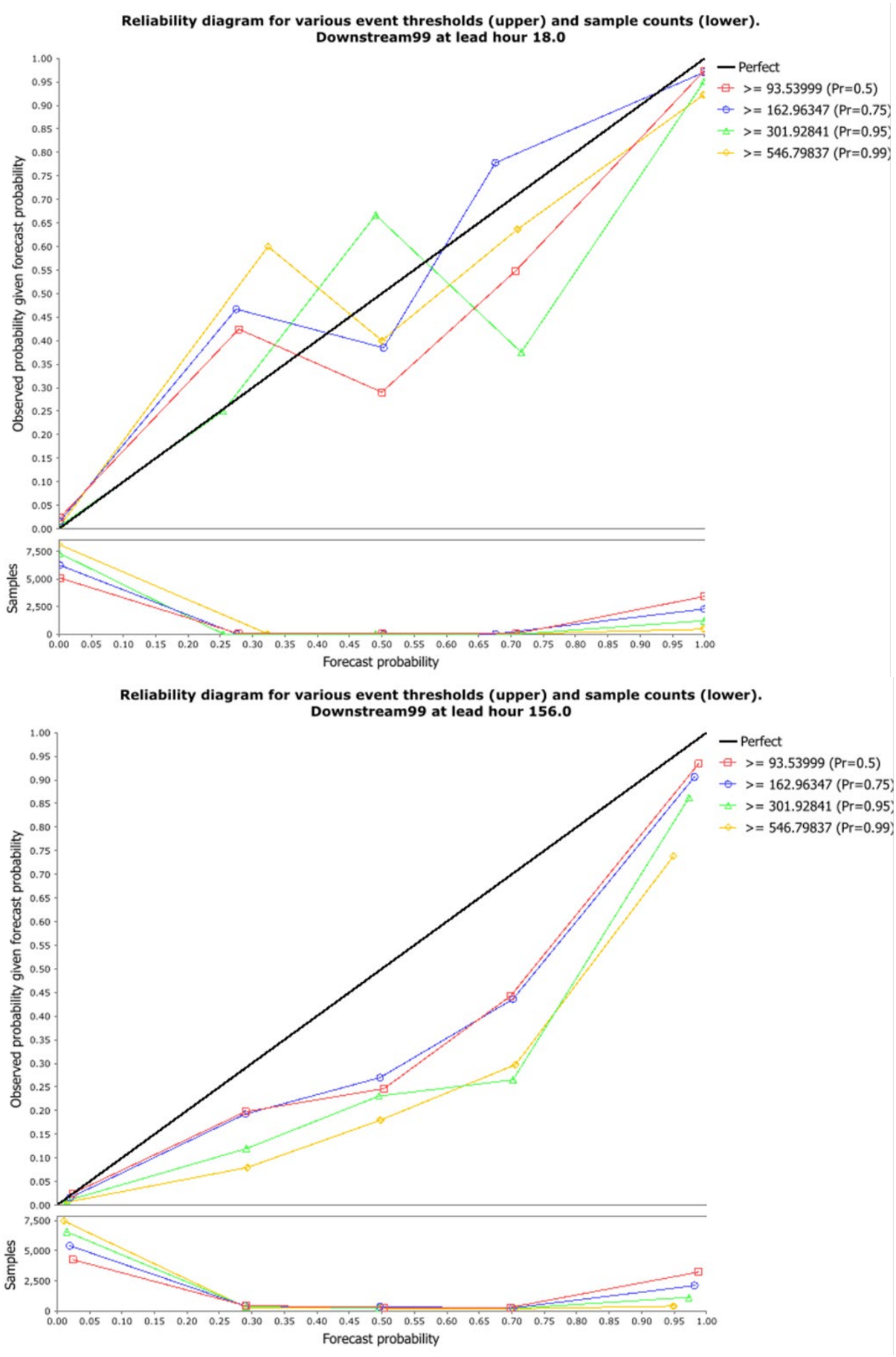
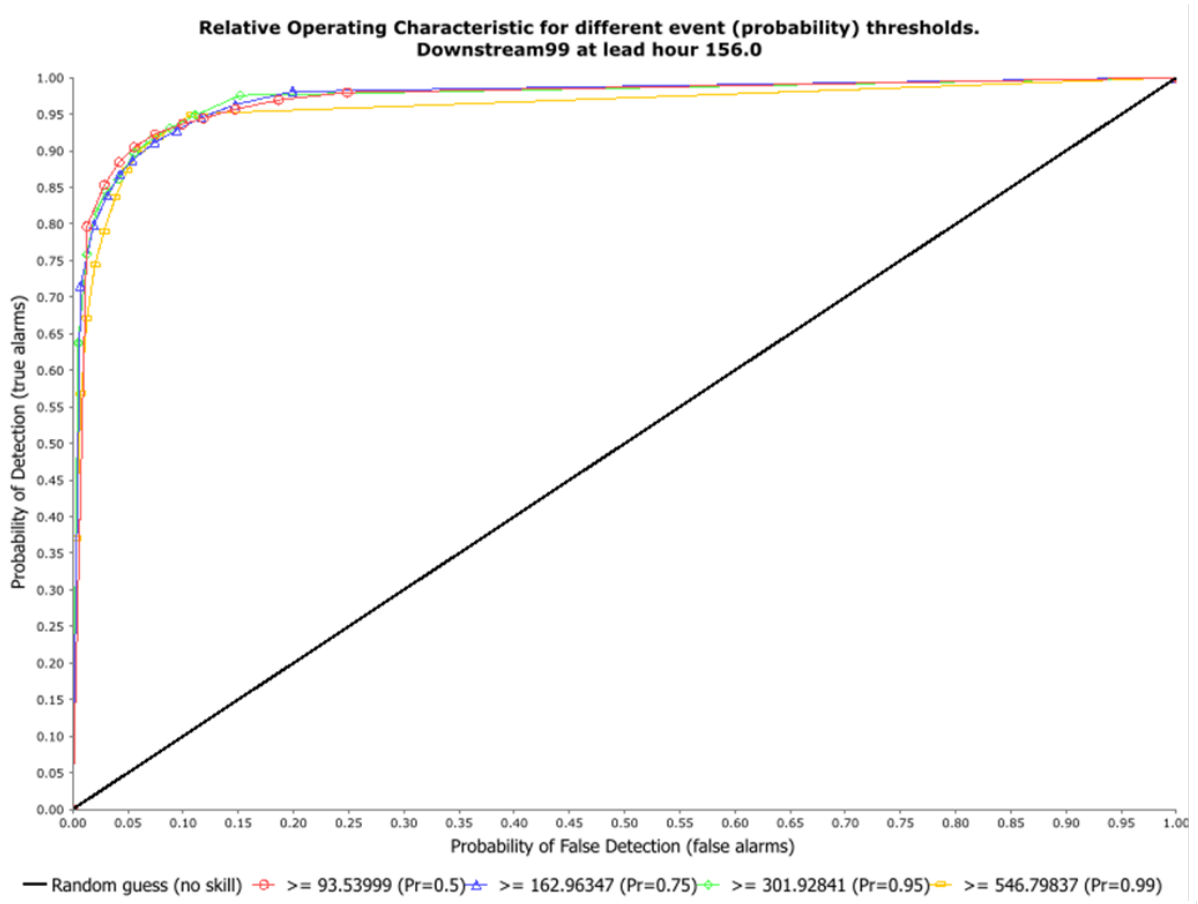


Figure 44. Reliability diagrams of NAEFS-forced ensemble streamflow forecast at lead times of 18 (top) and 156 (bottom) hours for downstream locations



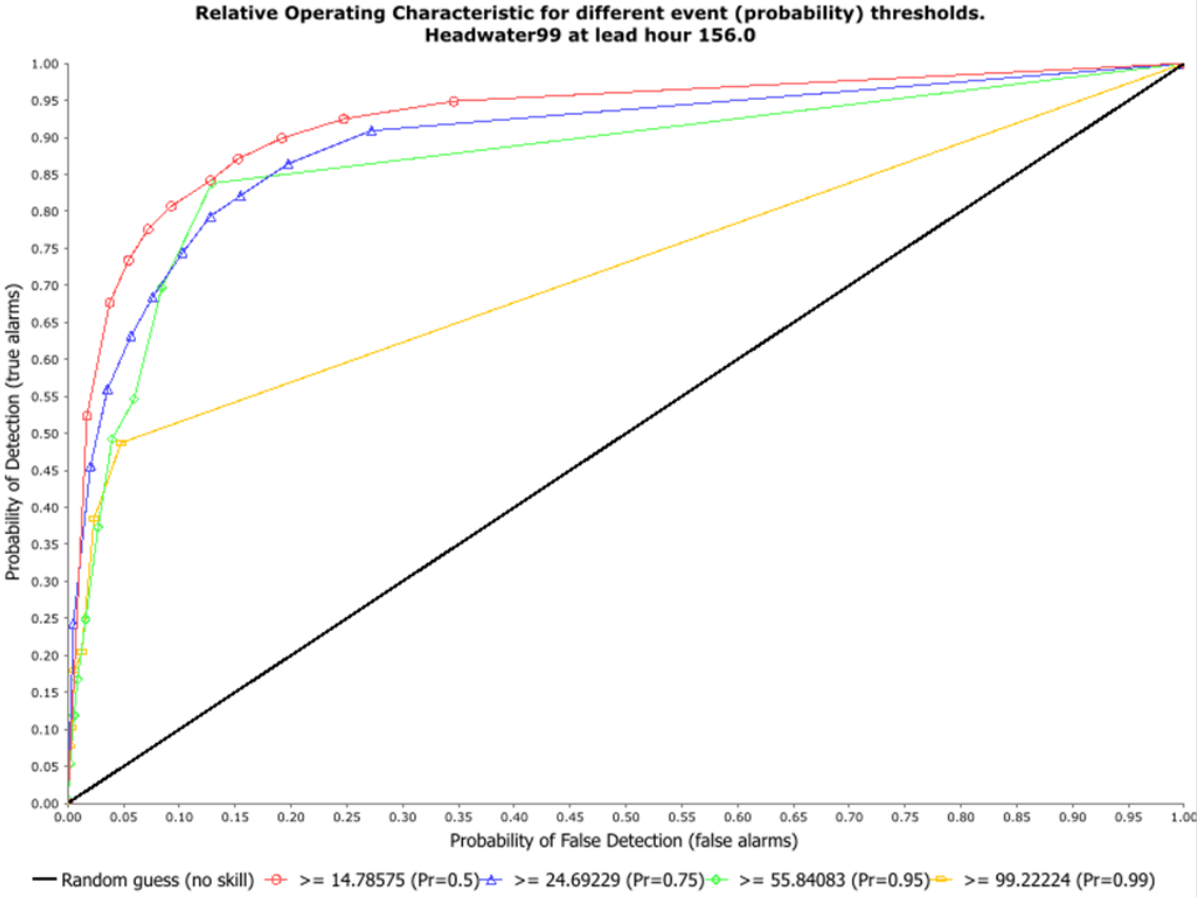


Figure 45. ROC curves for the NAEFS-forced ensemble streamflow forecast at different thresholds for downstream (top) and headwater (bottom) locations

7.4 Case 4. Verification of skill in ensemble streamflow forecast for water supply

Example 3 in Appendix A provides a hands-on exercise with a smaller version of this case using the EVS. Case 4 is adopted from Kim et al. (2018) and assesses the skill of ensemble streamflow forecasts generated with the HEFS (Demargne et al., 2014) for five headwater basins in the Upper Trinity River Basin (UTRB) in North Texas, United States. Straddling semi-arid and humid regions to the west and east, respectively, the UTRB is a challenging area for streamflow forecasting due to limited predictability and large input and hydrological uncertainties. The HEFS uses the MEFP (Schaake et al., 2007; Wu et al., 2011; NWS, 2017a) and the streamflow Ensemble Postprocessor (EnsPost) (NWS, 2017b) to reduce input and hydrological uncertainties, respectively (see Figure 1). For national implementation of the HEFS, it is of great interest to assess the possible gains from the use of bias-corrected medium-range precipitation forecasts from the MEFP and statistical bias correction in streamflow simulation via EnsPost in areas of limited predictability versus climatological ensemble forecasting.

Ensemble mean precipitation forecasts from the GEFSv10 (Zhou et al., 2017) were used as input to the MEFP to generate bias-corrected ensemble precipitation forecasts for lead times of 1 to 15 days. Beyond Day 15, resampled climatology (Demargne et al., 2014) was used – that is, skill-less ensemble precipitation forecasts generated with the MEFP (similar to climatological ensemble forecasts but based on MEFP ensemble generation). The period of record is 1985 to 2015. Even with the large-sample hindcast dataset, the sample size for large-to-extreme precipitation events is small for individual basins. To increase sample size, the forecast–observation pairs were pooled together for the five-basin cluster (see Kim et al., 2018 for the map).

Figure 46 shows the CRPSSs of the HEFS streamflow ensemble forecasts for aggregation periods of 1, 3, 5, 7, 14 and 30 days conditional on the verifying observation exceeding the respective 99th percentiles. The results are stratified for wet and dry seasons, and with and without EnsPost. For this case, only the conditional verification results are shown to accentuate the variations in skill and sampling uncertainty associated with large-to-extreme events. The CRPSS is expressed as:

$$CRPSS = 1 - \frac{\overline{CRPS}}{\overline{CRPS_{clim}}} \tag{50}$$

where $\overline{CRPS}$ and $\overline{CRPS_{clim}}$ denote the mean CRPS of the HEFS ensemble forecast and that of the resampled climatological forecast, respectively. NWS has been using climatological ensemble forecasting, referred to as ensemble (or "extended" when first implemented) streamflow prediction (ESP) (Day, 1985) for multiple decades for long-range streamflow forecasting. Hence, the resampled climatological ensemble forecast serves as a natural reference forecast.

Figure 46 shows that HEFS ensemble streamflow forecasts have discernible skill out to about 2 weeks into the future, and that EnsPost improves the skill by reducing hydrological uncertainty. Not surprisingly, the larger the time scale of aggregation is, the more skillful the forecast tends to be, an important consideration in forecast-informed operation and management of water supply systems in this water availability-sensitive region.

In all cases above, sample size is often a key limiting factor for verification of large-to-extreme events such as floods. Sampling uncertainty may be assessed via confidence intervals. For some verification measures, confidence intervals may be readily calculated using analytical expressions under assumed parametric distributions. Albeit computationally significantly more expensive, bootstrapping (see Chapter 2) may be used to estimate confidence intervals for any verification measure without distributional assumptions. Figure 47 shows the 90% (between 5% and 95%) confidence intervals estimated via Monte Carlo bootstrapping for the CRPSS for streamflow forecast for SGET2 (Clear Creek near Sanger, Texas, United States), one of the five forecast points used in Case 4. The figure indicates that the improvement due to EnsPost is statistically significant at a significance level of 0.10 except at very short lead times where

hydrological uncertainty is controlled largely by the memory of the ICs. The raw and postprocessed streamflow hindcasts are very likely to be significantly correlated with each other. It is hence very likely that the differences in the CRPSS between the two hindcasts are statistically more significant than the confidence intervals in the figure may suggest.

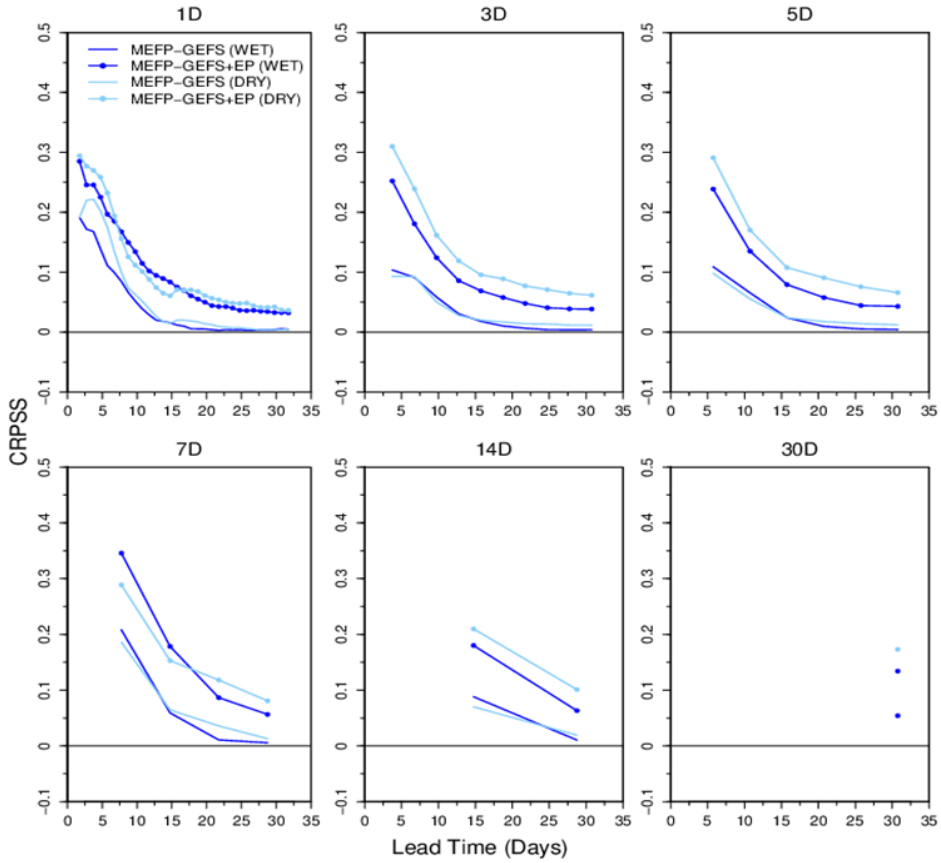


Figure 46. CRPSSs of the MEFP–GEFS streamflow ensemble forecasts with and without EnsPost (denoted as EP in the legend) for all aggregation periods for the wet and dry seasons conditional on the verifying observation exceeding the 99th percentile (from Kim et al., 2018)

Source: Reproduced with the permission of the American Meteorological Society

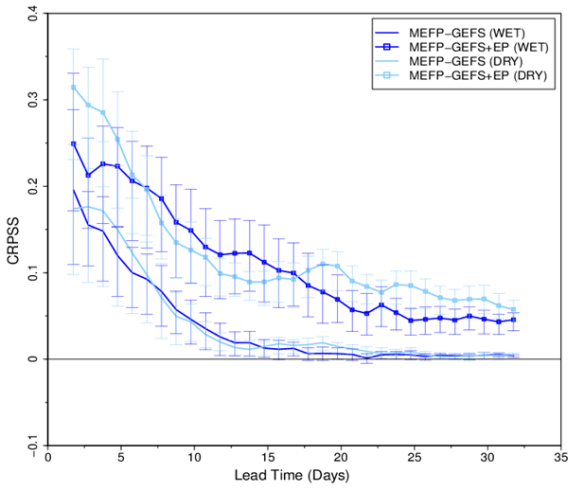


Figure 47. The 90% (between 5% and 95%) Monte Carlo confidence intervals for CRPSSs of daily streamflow hindcasts for SGET2 conditional on the verifying observation exceeding the 99th percentile (from Kim et al., 2018)

Source: Reproduced with the permission of the American Meteorological Society

7.5 Case 5: Diagnostic verification in (near) real time

Various hydrological forecasting systems include routines or subsystems for diagnostic verification in (near) real time. These measures are often computed or produced on a fixed schedule (for example, once a day or once a week). In the present section, this is referred to as operational verification.

Reasons for such operational verification include:

- Providing immediate feedback (as opposed to delayed feedback) to operational forecasters.
- Tracking of forecast quality for early identification of any issues.
- Enabling quick post-event analyses.

This section discusses three operational verification systems: the Système de Prévision Hydrologique (SPH) at the Québec Ministry of the Environment, Canada, the Hydrological Forecasting System (HyFS) at the Australian Bureau of Meteorology, and the Rijkswaterstaat Operational System (RWsOS) at the Water Management Centre of the Netherlands (WMCN). The purpose of this case study is to showcase various design considerations.

All systems are based on the Delft-FEWS forecast production system, and some metrics are computed using the EVS, which is used as a module within Delft-FEWS.

7.5.1 Système de Prévision Hydrologique (SPH)

The SPH is a Delft-FEWS-based hydrological forecasting system used by the Québéc Ministry of the Environment[1] to forecast streamflow for hundreds of locations in the province of Québec in Canada. Forecasts are used to inform other government agencies about hydrological conditions that may affect public safety. The system produces hydrological forecasts based on various meteorological forecast products which are referred to as forecast scenarios. These scenarios include blends of regional and global NWP products, including ensemble products: the National Centers for Environmental Prediction (NCEP) North American Mesoscale Forecast System (NAM) blended with the NCEP Global Forecast System (GFS), the Environment and Climate Change Canada (ECCC) Regional Deterministic Prediction System (RDPS) blended with the ECCC Global Deterministic Prediction System (GDPS), the NCEP Global Ensemble Forecast System (GEFS) and the ECCC Global Ensemble Prediction System (GEPS). A forecaster may modify these automatically produced forecasts prior to elevating the status of any of the scenarios to "official forecast".

The verification module within the SPH was developed to inform operational forecasters on the quality of recent streamflow forecasts and meteorological forecast scenarios. It includes a module that computes the relative mean error (RME) of streamflow forecasts and the mean error (ME) of precipitation and temperature forecasts (see Chapter 4 for definitions of RME and ME). For scenarios that comprise ensemble forecasts, the RME and ME of the ensemble means are computed. Computation is done as soon as verifying observations become available to the system. The computed metrics can be consulted within the same displays that forecasters typically use to produce the forecasts (that is, within the actual forecasting system). An example of the verification display is shown in Figure 47. The metric values are colour-coded and are presented as a function of the analysis time and forecast lead time.

[1] Its full name currently is the Ministry of the Environment, the Fight Against Climate Change, Wildlife and Parks.

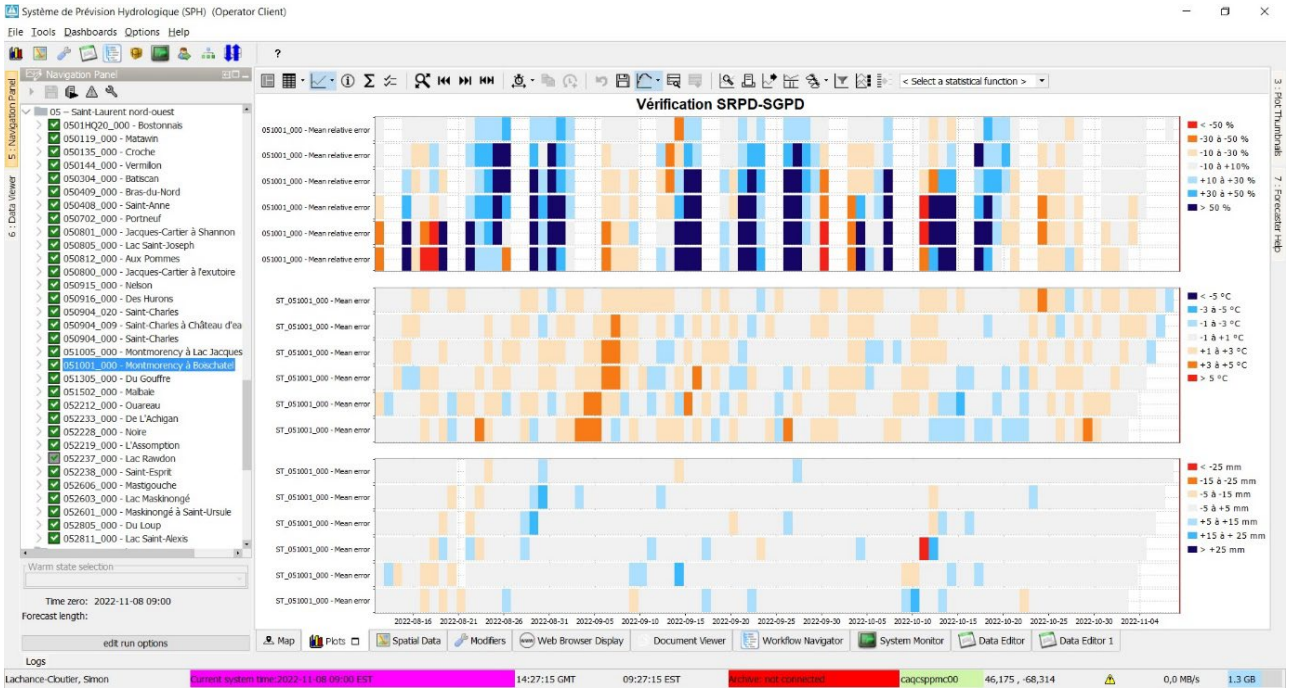


Figure 48. Example of the verification display used within the SPH. The display relates to a single location and its upstream basin. From top to bottom, the plots show the RME of the streamflow forecasts and the ME of precipitation and temperature forecasts. Within the plots, rows are for different lead times (not shown in the plot itself) and columns are for the analysis (initialization) times.

A main goal of the verification system is to provide operational forecasters with immediate feedback on the performance of recent forecasts. Such a feedback loop will increase the value added by human forecasters and hence improve forecast skill. The assumption here is that immediate feedback is better than delayed feedback.

Initial findings from the use of the system, however, indicate that forecasters find it difficult to use the verification metrics in their daily forecasting tasks. The quality of NWP forecasts, as expressed in the value of the ME, is not necessarily persistent over time, and hence the predictive value of recent verification metrics may be limited. Whereas human forecasters clearly add value to quantitative precipitation forecasts (QPFs) (Reynolds, 2003), it is not yet clear how immediate feedback may translate into improved forecast skill. Reynolds (2003) points out that, for human forecasters to add significant skill to QPFs, understanding of the physical processes and of the strengths and weaknesses of the NWP guidance is necessary.

7.5.2 Bureau of Meteorology's Performance Analysis Tool

The Hydrological Forecasting System or HyFS is the hydrological forecasting system used by the Australian Bureau of Meteorology. Service level agreements with emergency management partners for each Australian state and territory specify performance targets. Performance analysis compares flood watches, warnings and forecasts against key targets defined for over 500 forecast locations across Australia.

Performance is measured in three aspects: timeliness, warning lead time and forecast accuracy. The timeliness performance aspect aims to issue 97% of flood watches and flood warnings to customers on time (that is, before or at the stated "next issue time" in a previous watch or warning). The warning lead time target specifies that 70% of the trigger height exceedances for which a target lead time was provided to customers as per service level agreement are communicated within that target lead time. The forecast accuracy target specifies that 70% of the peak predictions provided to customers are within a specified range (typically ±0.3 m), as per service level specifications.

The flood forecasts and warnings issued by the Bureau are also available in machine readable .xml format. The .xml files containing the forecasts and warnings are imported into HyFS. HyFS runs a monthly workflow that calculates the performance metrics.

Within HyFS, various displays are available for displaying information related to the Performance Analysis Tool. The warnings display (Figure 49) allows the user to interrogate the system regarding each of the warnings issued and to undertake quality control. This includes exploring the forecasts underlying the warnings (Figure 50 and Figure 51). Summary data are available at the level of region, basin and location. From these plots (Figure 52 through Figure 54), it can be immediately determined whether the targets have been reached or not.

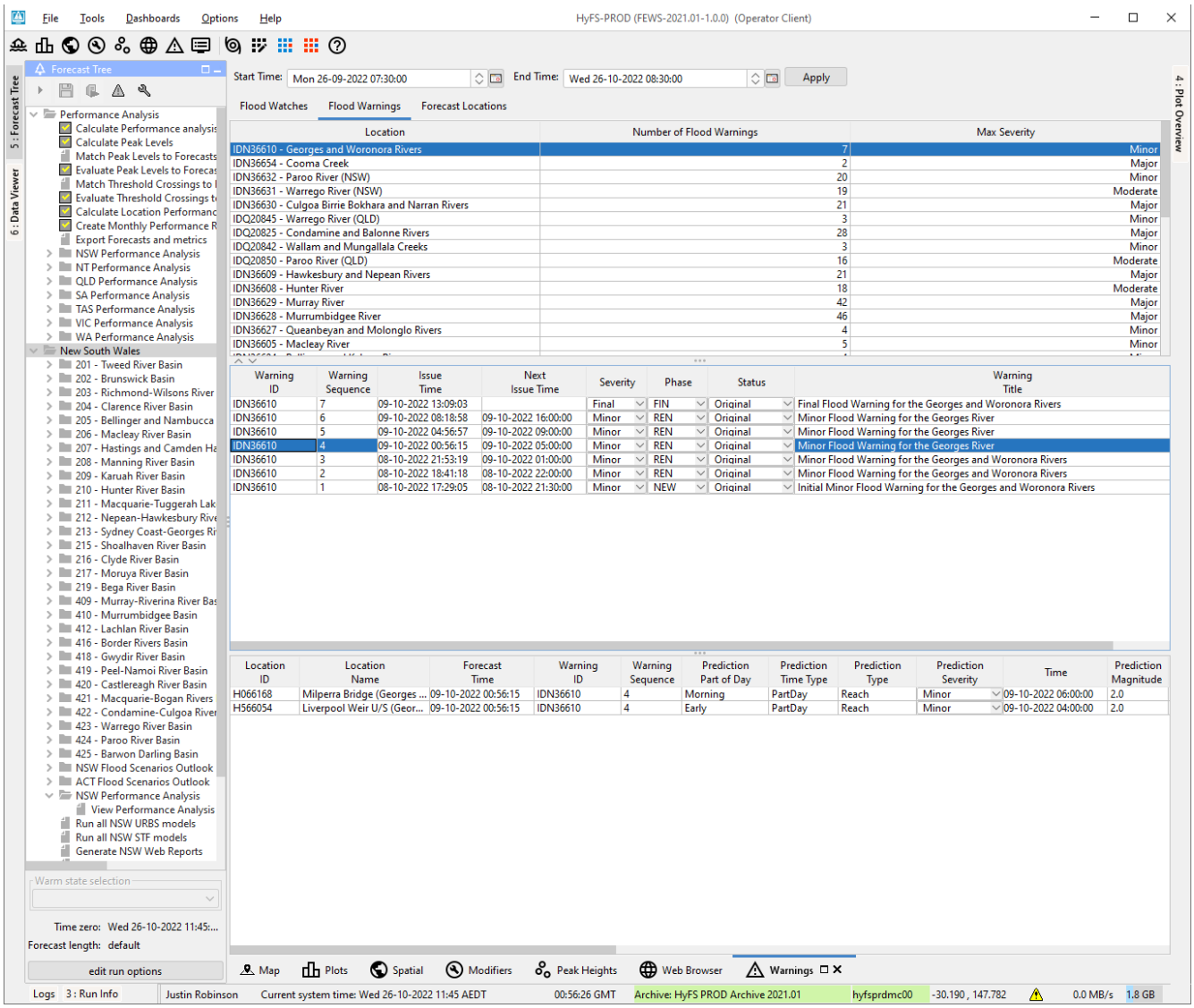


Figure 49. the HyFS warnings display

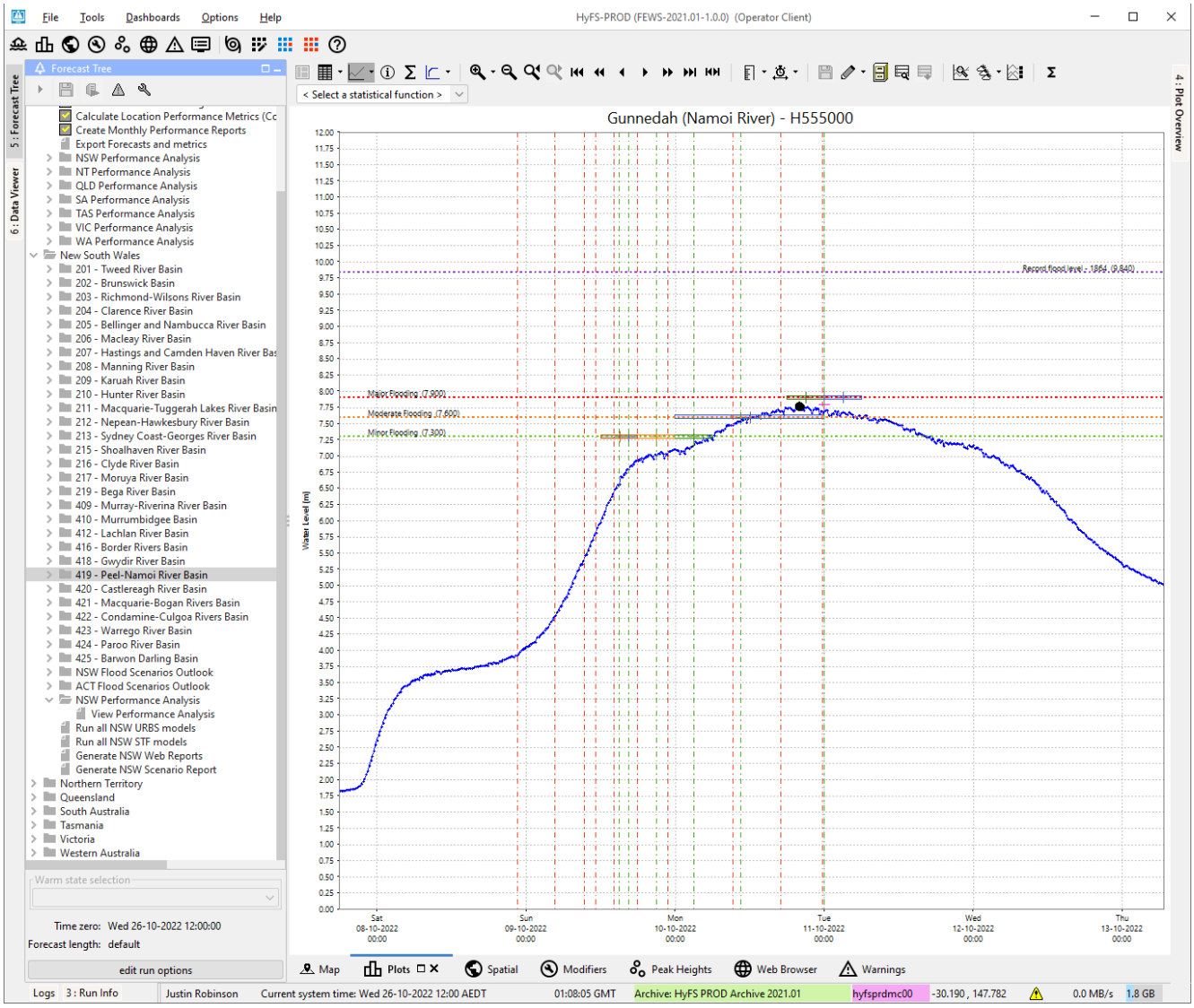


Figure 50. Forecasts underlying a specific warning, as shown in the HyFS warnings display

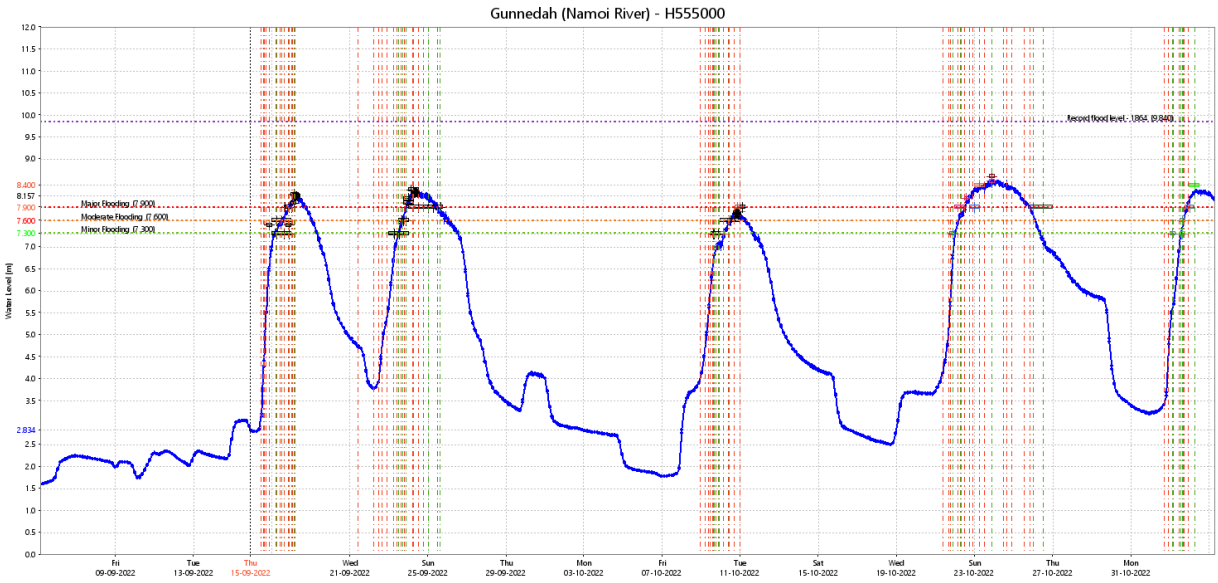


Figure 51. Detailed overview in HyFSof observed water level at a location, as well as the forecasts that were issued for that location. The latter are indicated by horizontal boxes that indicate a time interval.

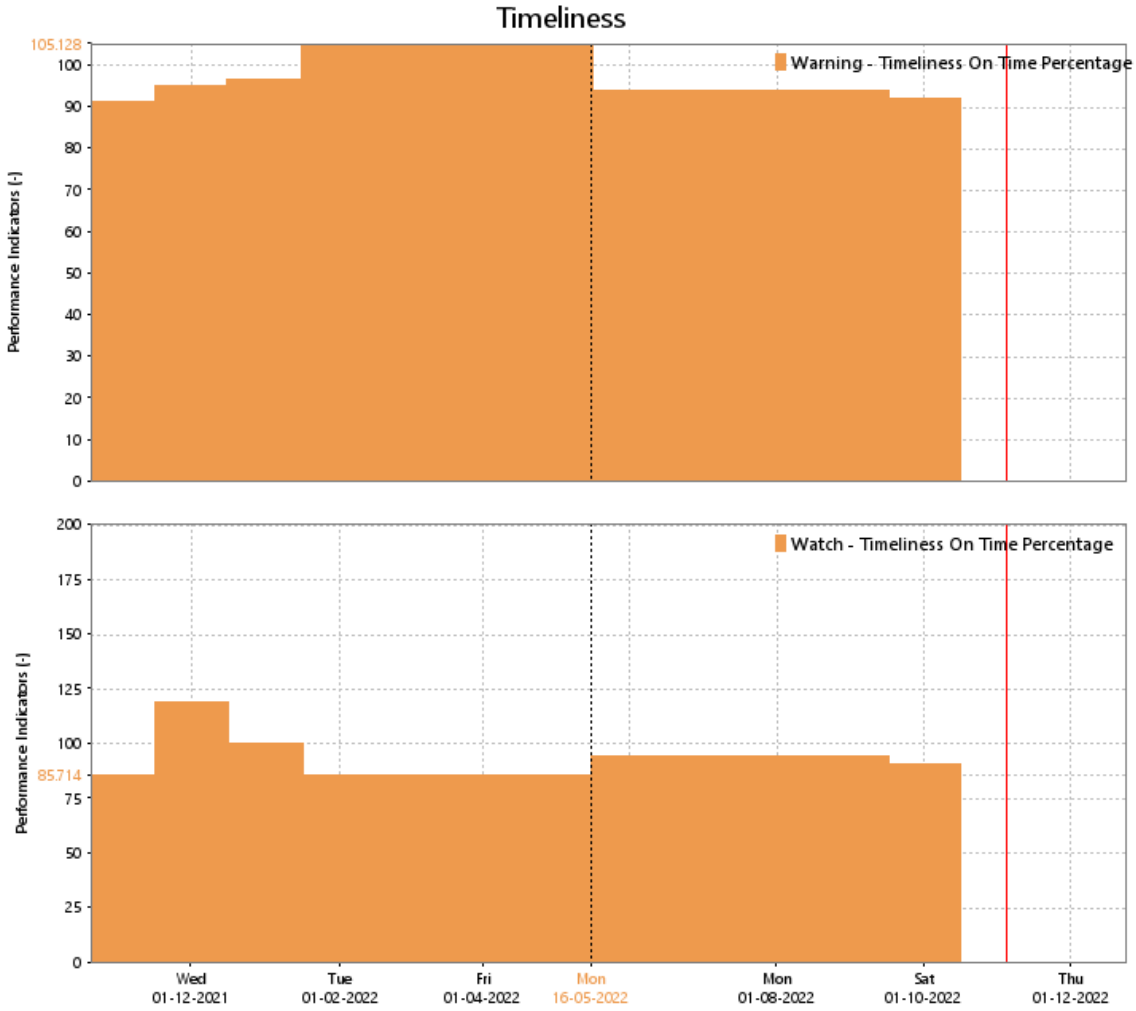


Figure 52. Timeliness display within HyFS

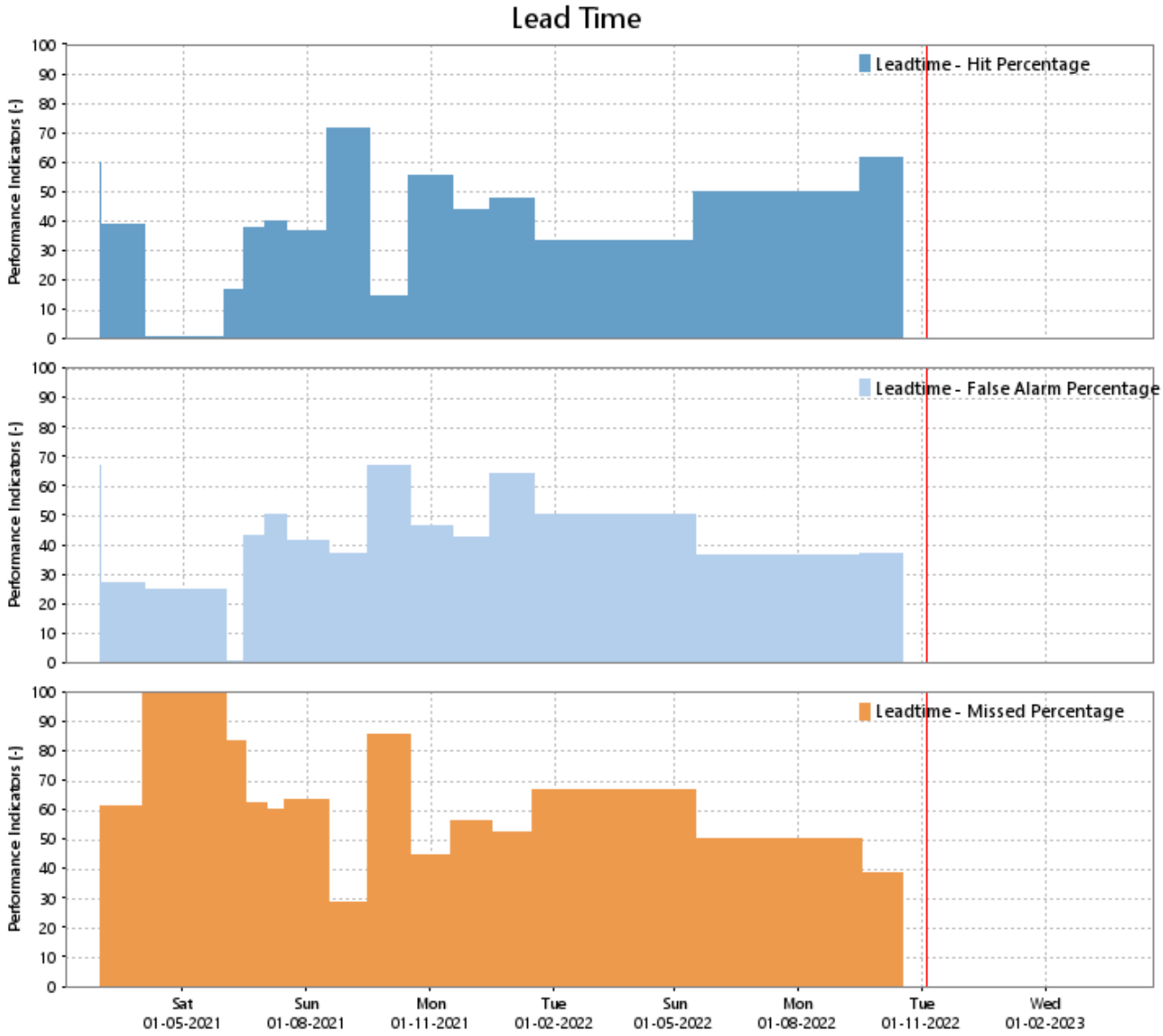


Figure 53. Lead time display within HyFS

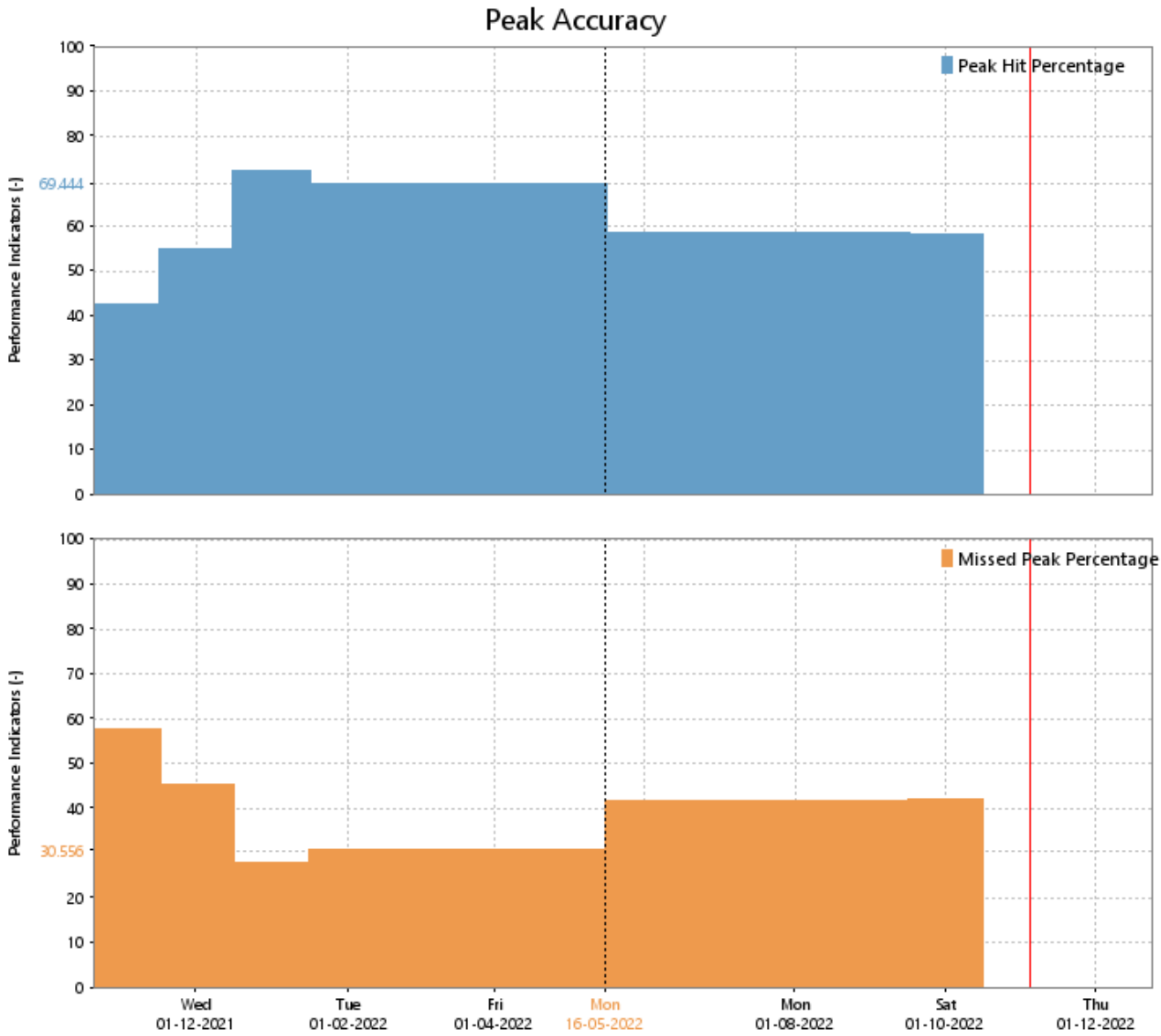


Figure 54. Forecast accuracy display within HyFS

Performance metrics, associated peaks and other information can be exported from HyFS. This facilitates detailed, offline analyses. An example of such an analysis is shown in Figure 55.

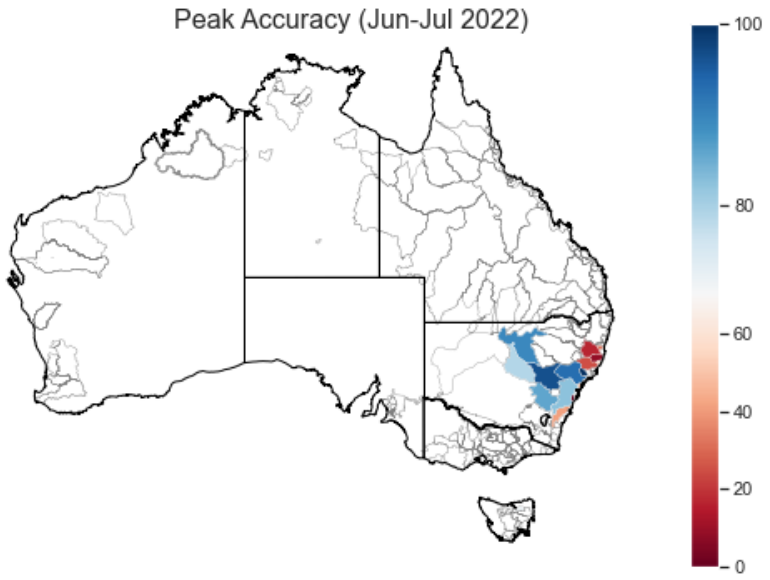


Figure 55. Peak accuracy, 16 June through 15 July 2022, by river basin. This map was produced using data exported from the Performance Analysis Tool in HyFS.

The performance metrics are used in monthly performance reports to provide feedback to forecasting teams. These reports contain a deeper analysis of the performance results nationally, while also reporting on forecasts for specific locations. They aim to create awareness of areas where the Bureau needs to improve its operations and allow operational forecasters to share their lessons learned with others. Also, they provide an opportunity to assess and celebrate improvements in the Bureau's service delivery. The Performance Analysis Tool has reduced the effort required to produce the performance reports, which allows for producing performance analyses more readily during (rather than after) events.

7.5.3 Rijkswaterstaat family of operational forecasting systems (RWsOS)

Rijkswaterstaat is an agency of the Government of the Kingdom of the Netherlands responsible for flood forecasting on water bodies that are under the purview of the central Government. Rijkswaterstaat operates hydrological forecasting systems for several types of water bodies, including the North Sea and the rivers Meuse and Rhine.

A verification system is being set up for the RWsOS family of forecasting systems. At the time of writing, the initial version has been in use for over a year. The verification system aims to enable early detection of any issues in the forecasting system and to provide the ability to give immediate feedback to operational forecasters. Issues include "drift", where model performance slowly deteriorates over time, and deterioration of the quality of initial model states.

The forecaster feedback mechanism will be developed in the coming years. This development will likely include research into how the feedback mechanism can lead to increased forecast skill, and training for operational forecasters on how to achieve that.

Currently, the verification tool is being expanded to include verification metrics computed by the Ensemble Verification System (EVS). The EVS will be included as a module within the Delft-FEWS based RWsOS products. The RWsOS system is used to source data required for verification. These data are exported in the Delft-FEWS Published Interface format, which can be readily ingested by the EVS. An EVS-Delft-FEWS "adapter" prepares the EVS project file which is required by the EVS to run the computations. Once these have been completed, a "post-adapter" casts the EVS outputs into a format that can be ingested by the archive which is linked to Delft-FEWS. Information is then available for visualization and for further dissemination to a web portal (the "verification dashboard") which was developed for exploring the verification data.

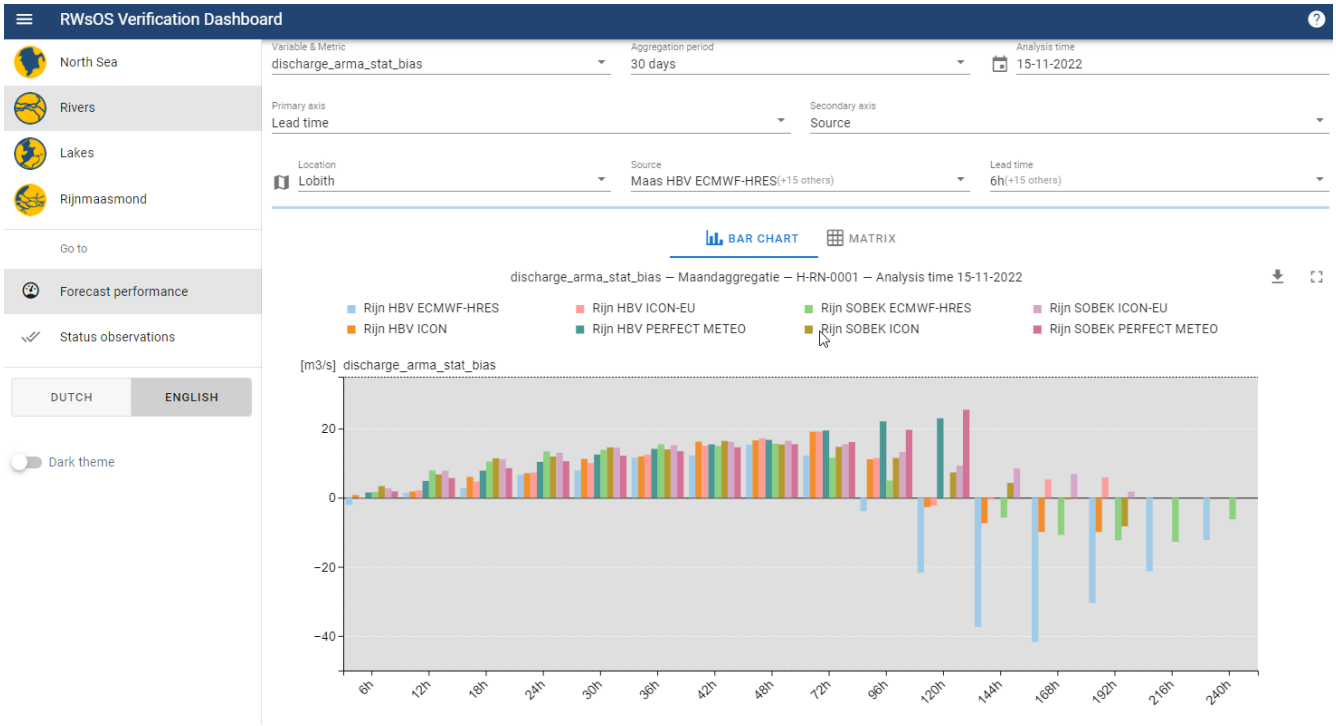


Figure 56. Screenshot of the Rijkswaterstaat verification dashboard

7.6 Case 6: Comparative verification of ensemble forecasts for ephemeral streams

Verification of ensemble forecasts for ephemeral streams tends to be unusual in the wider literature. Example 7 in Appendix A is a real-world example of an attempt to verify ensemble streamflow forecasts of an ephemeral river and provides methods and code to compare two ensemble streamflow forecasting systems, with selected forecasts and data from the study by Bennett et al. (2021). The full set of forecasts and observed data are provided by Bennett and Robertson (2021).

The forecasts are generated by forcing a calibrated and initialized rainfall-runoff model with "perfect" rainfall forecasts (observed rainfall), and uncertainty is generated with two different versions of the Error Reduction and Representation in Stages (ERRIS) error model (Li et al., 2017). ERRIS transforms, bias-corrects and updates forecast errors, and then propagates uncertainty through a technique termed "stochastic updating". ERRIS is expressly designed to handle ephemeral streams with zero flow.

Example 7 in Appendix A compares a "new" version of ERRIS to an "old" version. The old ERRIS uses a static bias-correction and applies a restriction (Li et al., 2015) to the autoregressive updating during parameter inference and forecast generation. This restriction attempts to avoid unrealistically large corrections. The new ERRIS uses a moving average bias-correction and does not apply the restriction to autoregressive updating when parameters are inferred but does apply the restriction in forecast generation. For detailed description of the ERRIS models used to generate these forecasts see Bennett et al. (2021).

Typically, when comparing forecasting methods, the choice of the method depends on the purpose of the system. In this case study, the old method produces slightly more accurate and reliable forecasts at longer lead times (>~150 hours) but tends to be positively biased. The new method is more accurate and reliable at shorter lead times (<~150 hours) and is largely unbiased. At longer lead times the new method becomes overconfident, reducing reliability.

The use of pseudo-PIT values is extremely useful in assessing reliability of forecasts for ephemeral streams as it allows us to check the uniformity of PIT values and associated

summary methods (α-index, β-score, β-bias) (see Example 7 in Appendix A). It is noted, however, that when there are a very large number of zero flows observed, the uniformity of PIT will largely rely on pseudo-PIT values. This can be difficult to interpret: in such cases, reliability essentially comes down to the probability of zero drawn from each ensemble forecast.

7.7 Key points

- The case studies demonstrate how relatively simple verification tasks may aid critical operational and strategic decisions for improving forecast quality.

- The case studies show how verification functionalities may be integrated with real-time forecast or other operational systems to very quickly provide the forecasters, system operators or managers with forecast quality information specific to recent events or periods as desired.

- Verification of streamflow forecasts tends to make use of a few metrics, scores and diagrams out of a large number available. They comprise representative accuracy metrics and attribute-specific metrics that are largely independent of one another in information content.

- For single-valued (deterministic) streamflow forecasts, the core metrics tend to include the RMSE, ME and correlation. For probabilistic streamflow forecasts, the core metrics tend to include mean CRPS, reliability diagram (if sample size is insufficient, rank histogram or PIT diagram), forecast frequency histogram and ROC curve.

- With the availability of software tools for verification, it is easy to generate numerous verification metrics, scores and diagrams. Unless purposefully sought, however, the verification results may not translate into actionable information or offer insight by themselves. It is hence important to develop a verification plan even if it may have to be revised. For example, one may be able to show that forecast A has a higher skill score than forecast B, but it may be difficult to assess the practical value of adopting forecast B in place of, or in addition to, forecast A without assessing the marginal value in physical-world terms, even only very approximately (see Murphy and Ehrendorfer (1987) and Laugesen et al. (2023)).

- The quality of streamflow forecasts tends to be flow regime-dependent. For example, forecast A may be better than forecast B when all ranges of flow are considered, but in high flow conditions, forecast B may be superior. To assess such flow magnitude-dependent variations of forecast quality, conditional verification should also be used as necessary. Because high-flow events are generally more impactful, such additional verification information, beyond the "unconditional" verification information, is very important for many users of the forecast information.

- Planning is particularly important for verification of large-to-extreme events for which a large sample is generally necessary. If a suitable dataset does not exist, reforecasting may be necessary, which requires substantial organizational resources and commitment. Note that, even if large-sample reforecasts of hydrometeorological variables are already available, large-sample hydrological hindcasting for many locations requires significant human and computational resources.

CHAPTER 8. SUMMARY

This chapter compiles the key points from Chapters 2 through 7. Because they appear in the general order of initiating hydrological verification, this chapter may be read as the overall guidance without the details. As no two rivers are the same and no two users are the same, no two streamflow verification tasks are likely to be the same. To address the specific verification objectives and needs effectively and time-efficiently, it is important to gain understanding of and familiarity with the fundamental concepts, the available metrics, scores, diagrams and tools, and their limitations.

8.1 Chapter 2: Why verify?

- The purpose of hydrological verification is to increase the value and utility of hydrological forecast products and services by supporting objective and systematic improvement of forecast quality and the decisions of the users of the forecast information.

- Hydrological verification broadly utilizes the theory and practices developed by the meteorological community, which was early to recognize the value of verification in improving weather forecasting.

- Hydrological forecasts are subject to input and hydrological uncertainties. The former are associated with errors in the hydrometeorological forecasts used as input to hydrological models. The latter are associated with errors in the rest of the hydrological forecasting process. Verification supports uncertainty decomposition to guide cost-effective improvement of forecast input, systems and processes.

- Type I, or false positive, and type II, or false negative, errors are competing attributes of a forecast. Verification informs the trade-off between the two types of errors and supports decision-specific assessment of the utility of a forecast and the relative utility of competing forecasts.

- Movement and storage of water is heavily modulated by the physiography of the individual catchments, river basins, channels and water bodies. Hence, it is generally not possible to determine streamflow at other locations based on observations at a gauged location. For this reason, hydrological forecasts should be verified as location-specifically as possible to the extent data availability allows. Note that an "accurate" precipitation forecast at a regional scale to a meteorologist can very easily be a complete miss to a hydrologist or a water resources engineer particular for small basins if precipitation falls outside of the boundary of the catchment of interest.

- Prediction of large-to-extreme events is very often the most important service of operational hydrological forecasting. Such events occur infrequently, and hence the sample size tends to be small. To increase sample size, some form of trading of space for time or regionalization is usually necessary at the expense of location specificity.

8.2 Chapter 3: Attributes of forecast quality

- Forecast quality is assessed by comparing forecasts with verifying observations (or high-quality estimates) under the assumption that they are realizations of IID random variables. The relationship between the two is then described wholly by their joint distribution. Multiple attributes of forecast quality are necessary to describe the essence of this distribution.

- Accuracy describes the overall level of agreement between the forecasts and their verifying observations and hence is most representative of forecast quality. Measures of

accuracy, such as the RMSE and mean CRPS for single-valued and probabilistic forecasts, respectively, reflect multiple attributes that are largely independent of one another in information content, such as correlation and biases in the mean and standard deviation in the case of the RMSE.

- Skill describes the relative accuracy of the subject forecast in comparison with a reference forecast or benchmark of choice. The reference forecast may be climatology, persistence or a forecast produced from a baseline forecast system. Skill scores calculate percent improvement in the accuracy metrics of choice by the subject forecast over the reference forecast. The Nash–Sutcliffe efficiency, which is very widely used in calibration of hydrological models, is an example of the MSE skill score.

- Several attributes of probabilistic forecasts arise from decomposing the joint probability distribution of the forecast and the verifying observation into the conditional and marginal distributions.

- In probabilistic verification, reliability (or type I conditional bias), resolution and uncertainty arise from conditioning on the forecast via the calibration–refinement (CR) decomposition, whereas type II conditional bias, discrimination and sharpness arise from conditioning on the observation via the likelihood–base rate (LBR) decomposition.

- Though verification and prediction are different, it is helpful to relate the CR decomposition with forward (that is, regular) linear regression and the LBR decomposition with reverse regression (that is, regression with the predictor and the predictand interchanged). It is also helpful to consider the CR and LBR decompositions as characterizing forecast quality from the perspective of reducing false alarms (crying wolf when there are none) and misses (failing to see the wolf), respectively, given the same absolute accuracy in the forecast.

- Reliability, resolution, type II conditional bias and discrimination are competing attributes given a particular absolute accuracy in the forecast. Specifically, reducing type I and type II conditional biases is a zero-sum game unless absolute accuracy is improved. Hence, assessment of individual forecast attributes is critical to assessing the trade-offs, guiding improvements in forecast systems and processes, and improving application-specific decisions based on the user's risk perception and tolerance.

- In the CR decomposition, uncertainty reflects predictability of the variable being verified. In the LBR decomposition, sharpness measures the forecast's ability to "stick its neck out", correctly or incorrectly. Though these two attributes do not pertain to the joint relationship between forecast and observation, they make up the overall accuracy and hence should be assessed. When assessing uncertainty, it is a good practice to consider skewness (asymmetry in distribution) and heteroscedasticity (nonuniformity in variability) to aid possible stratification, pooling or conditioning of the forecast–observation pairs.

- The above points regarding probabilistic verification mean that, between reliability and resolution and between type II conditional bias and discrimination, it is generally necessary to assess only one of the two attributes in each pair. Commonly, the choices are reliability and discrimination.

8.3 Chapter 4: Commonly used verification metrics

- Multiple metrics are available to assess each of the widely measured attributes for different types of forecasts. The MSE (or RMSE), the 2 × 2 contingency table, the BS and the mean CRPS are particularly important, as they collectively contain almost all other metrics or their building blocks. It is hence important to understand what each of the above four metrics comprise, so that one may utilize the entire suite of metrics, scores and diagrams effectively.

- This publication describes most of the widely used verification metrics, scores and diagrams. Some of the metrics derived from the 2 × 2 contingency table are better suited for verification of flash flood forecasts than streamflow forecasts. All others described in this chapter apply to the verification of streamflow forecasts. Most of them are available in verification software tools such as the EVS.

- Several diagrams and histograms are used to assess the attributes associated with the BS decomposition for probabilistic forecasts. The most widely used include the reliability diagram (a strong test of reliability), the rank histogram (a weak test of reliability), the ROC curve for discrimination and the forecast frequency histogram for sharpness. Depending on the application, the PIT diagram (see Case 6 in Chapter 7) and discrimination diagram may be preferred to the rank histogram and the ROC curve, respectively.

- Streamflow and precipitation typically exhibit skewness (asymmetry in distribution) and heteroscedasticity (nonuniformity in variability), which are often not explicitly assessed in verification. For streamflow, skewness and heteroscedasticity reflect predictability and flow regime-dependent variability, respectively, and hence provide very useful guidance on stratification, pooling, or conditioning of the forecast–observation pairs.

- Sampling uncertainty is a recurring challenge in hydrological verification, particularly for conditional verification of probabilistic forecasts for large-to-extreme events. It is a good practice to assess sampling uncertainty via bootstrapping early in the verification task for sample size-challenged cases. One may then assess how sampling uncertainty may be reduced by relaxing the conditioning or trading space for time via data pooling or regionalization.

- Conditional verification results, if produced, should be communicated together with the "parent" unconditional verification results to avoid misinterpretation or misuse. For example, using verification information for high flows as being representative of all flows will inevitably lead to poor decisions most of the time.

8.4 Chapter 5: Preparatory steps and logistical considerations

- Several steps are generally necessary to produce useful verification results. Each step involves several choices and decisions which, if not well thought out in advance, may lead to avoidable trials and errors.

- The high-level steps include: defining the verification objectives, determining the audience, selecting verification tools, collecting data, preparing data (including pairing), and computing verification statistics.

- The verification process may vary greatly depending on the verification task. It is hence very important to clearly define the verification objectives before starting.

- The verification audience determines how verification information may be presented. Possible audiences include forecast users, decision support staff, forecasters, forecasting system developers and administrators.

- Verification is generally data intensive and hence software tools are almost always used. Care should be taken to select a tool or tools that will meet the verification objectives and fit the resources available to the organization and to those who will use them. In addition, the tools should be supported, maintained and updated well into the future.

- The data required for verification may be readily available or may have to be produced through hindcasting (that is, reforecasting), which has potentially significant resource implications.

- The collected or generated data may have to be cast into a specific format for ingestion by the verification tool. Forecasts and observations must be paired for verification. Depending on the verification task, stratification, subsampling or conditioning of the full dataset may be necessary.

- Once the forecasts and observations are paired and subsampled as necessary, verification metrics may be computed. Several verification measures are expressed as diagrams with standardized layouts. Often, the verification task requires uncertainty bounds to be computed.

8.5 Chapter 6: Visualization of verification information

- Verification information may include raw data, technical diagrams, summary metrics and metadata.

- Verification information generally includes a large amount of metadata.

- Verification can yield a large number of verification statistics. These cannot all be shown in a single plot or table, and hence accompanying metadata will have to identify the "dimensions" that are not explicitly communicated in the plot or table.

8.6 Chapter 7: Case studies

- The case studies demonstrate how relatively simple verification tasks may aid critical operational and strategic decisions for improving forecast quality.

- The case studies show how verification functionalities may be integrated with real-time forecast or other operational systems to very quickly provide the forecasters, system operators or managers with forecast quality information specific to recent events or periods as desired.

- Verification of streamflow forecasts tends to make use of a few metrics, scores and diagrams out of a large number available. They comprise representative accuracy metrics and attribute-specific metrics that are largely independent of one another in information content.

- For single-valued (deterministic) streamflow forecasts, the core metrics tend to include the RMSE, ME and correlation. For probabilistic streamflow forecasts, the core metrics tend to include mean CRPS, reliability diagram (if sample size is insufficient, rank histogram or PIT diagram), forecast frequency histogram and ROC curve.

- With the availability of software tools for verification, it is easy to generate numerous verification metrics, scores and diagrams. Unless purposefully sought, however, the verification results may not translate into actionable information or offer insight by themselves. It is hence important to develop a verification plan even if it may have to be revised. For example, one may be able to show that forecast A has a higher skill score than forecast B, but it may be difficult to assess the practical value of adopting forecast B in place of, or in addition to, forecast A without assessing the marginal value in physical-world terms, even only very approximately.

- The quality of streamflow forecasts tends to be flow regime-dependent. For example, forecast A may be better than forecast B when all ranges of flow are considered, but in high flow conditions, forecast B may be superior. To assess such flow magnitude-dependent variations of forecast quality, conditional verification should also be used as necessary. Because high-flow events are generally more impactful, such additional verification information, beyond the "unconditional" verification information, is very important for many users of the forecast information.

- Planning is particularly important for verification of large-to-extreme events for which a large sample is generally necessary. If a suitable dataset does not exist, reforecasting may be necessary, which requires substantial organizational resources and commitment. Note that, even if large-sample reforecasts of hydrometeorological variables are already available, large-sample hydrological hindcasting for many locations requires significant human and computational resources.

Hydrological verification is still in its early days. As evidenced by several examples and case studies presented in this document, there are limits to the applicability of the current theory and practices. For example, verification of time-to-peak forecast is not addressed in this document despite its importance. Whereas one might be able to carry out probabilistic verification of ensemble time-to-peak forecast derived from ensemble streamflow forecast, such an attempt decouples phase from amplitude in a hydrograph, yielding verification results that are not amenable to hydrological contextualization. Such verification information is not likely to provide the producers of the forecast with tangible clues or insight into improving forecast quality or the users of the forecast with additional skill for improved decision-making. In this regard, both the research and the operational communities have much to contribute to the advancement of the science and practice of hydrological verification.

APPENDIX A. HANDS-ON EXAMPLES OF FORECAST VERIFICATION

This appendix describes various examples of verification of streamflow forecasts. The examples use the Ensemble Verification System (EVS) (Brown et al., 2010), the R verification package and the Python-based verif library.

The examples are meant to provide possible templates for a range of verification tasks (see Chapter 7) one may encounter when embarking on hydrological verification. Though unlikely within the foreseeable future, the coding examples may become outdated at some point due to upgrades of the underlying software. The expectation is that the hands-on examples will have served their useful purpose by then and new examples of hydrological verification will emerge that leverage the latest advances in science and technology.

A.1 Setting up the Ensemble Verification System (EVS) for Examples 1–3

The EVS package version 5.10 is available at https://sourceforge.net/projects/ensemble-verification-system/. For installation and start-up on Microsoft Windows or Linux, go to page 9 of the EVS Manual (included in the package) and follow the instructions. The files necessary for the examples are available at https://wmostorage.blob.core.windows.net/wmo-public/hydro-forecast-verification-guidelines/A1-3_EVS_example/A1-3_EVS_example.zip (Figure 57).

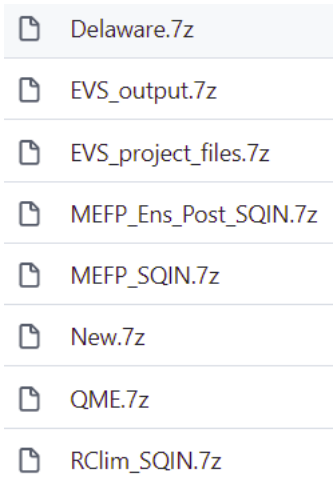


Figure 57. Files to be downloaded for Examples 1 through 3

In Figure 57, "New" happens to be the name of the river in Example 1 and has no literal significance. Once all eight files are downloaded, unzip them. This should result in the list of folders shown in Figure 58.

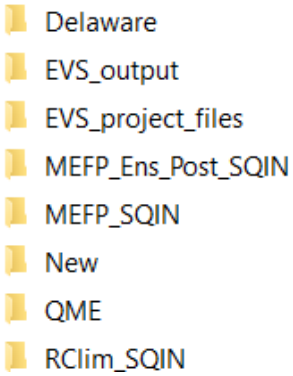


Figure 58. Folders for Examples 1 through 3 once the downloaded files are unzipped

If the user does not have 7-Zip, which is a free, open-source file archiver, it should be downloaded and installed from https://sourceforge.net/projects/sevenzip/. Create a new folder named UTrinity and move the four folders *MEFP_Ens_Post_SQIN*, *MEFP_SQIN*, *QME* and *RClim_SQIN* into it, which should result in the list of folders shown in Figure 59.

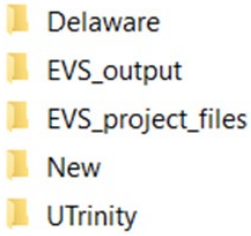


Figure 59. High-level folders for conducting verification of streamflow forecasts using the EVS for Examples 1 through 3

Lastly, unzip all files in the folders *Delaware* and *UTrinity*. In the examples below, it is assumed that the above folders are located under E:\verification\. **The user must modify all occurrences of the above path in the EVS project files and in this appendix to run the EVS and navigate correctly.**

The *New* folder contains all forecasts and observations for Case 1 in Chapter 7. The *Delaware* and *UTrinity* folders contain small subsets of the forecasts and observations for Cases 3 and 4 in Chapter 7, respectively. The *EVS_project_files* folder contains all EVS project files for the three examples. The *EVS_output folder* is initially empty and is to contain all output files from the EVS runs. The descriptions below assume that the user has at least scanned through the EVS Manual for general understanding of the software structure and the overall flow of the verification operation.

A.2 Example 1 – Verification of single-valued streamflow forecast with uncertainty decomposition

Navigate to *E:\verification\EVS_project_files\New* and double-click on *GEFSv12_*ens*_mean_SQIN*. The first part of the project file name, *GEFSv12_ens_mean*, refers to the forcing used (that is, ensemble mean precipitation forecast from GEFSv12). The second part of the project file name, *SQIN*, is United States National Weather Service (NWS) parlance for simulated (S) instantaneous (IN) discharge (Q), in which "simulated" means model-generated. The project file name hence indicates that this verification is for GEFSv12 ensemble mean-forced streamflow forecast.

Once the EVS is successfully launched, the graphical user interface should show the following three verification units (VU) in subpanel *1. Add verification unit(s)*:

1. *GAXV2.GEFSv12_ens_mean_fcst_flow* (for total uncertainty in Figure 35)

2. *GAXV2.GEFSv12_ens_mean_sim_flow_with_DA* (for hydrological uncertainty without the IC uncertainty in Figure 35)

3. *GAXV2.GEFSv12_ens_mean_sim_flow_without_DA* (for hydrological uncertainty in Figure 35)

For each VU, the locations of the forecast and observed data sources may be found under the subpanel *2b. Set input data sources*. The location of the output files may be found under the subpanel *2d. Set path for outputs*. As noted above, it is important that the user modify the *:E\verification* part of the paths for all three folders (forecast, observation and output) to match the actual locations of the folders on the user's computer.

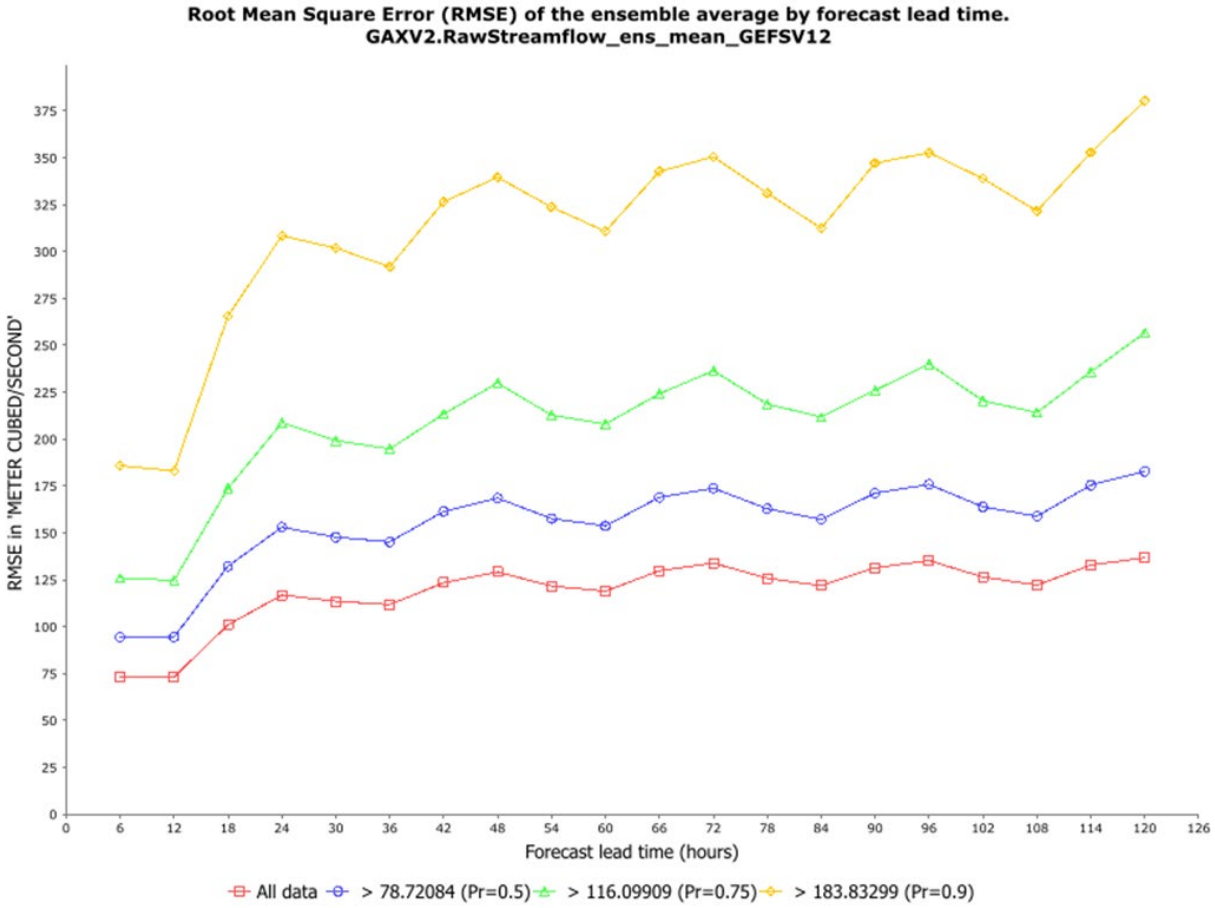


Figure 60. EVS-generated RMSE result for GAXV2 (New River at Galax, Virginia, United States)

Click on the *Next* button in the lower-right corner of the *Verification* panel. In the *Verification metrics to compute* panel, verify that *Sample size*, *Correlation coefficient*, *Mean error* and *Root mean square error* are checked. With these metrics calculated, one may construct the MSE decomposition result shown in Figure 36 in Chapter 7. To proceed, click on the *Run* button in the lower-left corner of the panel.

Once the above run is complete, the user may navigate to *E:\verification\EVS_output\New\GEFSv12_ens_mean_SQIN\GAXV2* and verify that two XML files ending with *pairs_raw* and *pairs_cond* have been generated (see page 35 of the EVS Manual for explanation). The two files are used to calculate the verification statistics. On the *Verification* panel, click on the *Next* button in the lower-right corner for the *Aggregation* panel. Example 1 involves only a single forecast point, and hence aggregation does not apply. Click on the *Next* button in the lower-right corner for the *Output* panel.

On the *Output* panel, click on *GAXV2.GEFSv12_ens_mean_fcst_flow* or on *VERIFICATION* in the same row in subpanel *1a. Select unit(s) with results* to list all possible products in subpanel *1b. Choose products for selected unit*. Right-click anywhere within the rectangular area of the list of products and choose the option *Select all times and products*, which will check all boxes. Click on any one of the products to display all lead times in subpanel *1c. Choose lead times for selected products* and verify that all boxes are checked. Click on the *Run* button in the lower-left corner of the panel to generate the products.

To see all verification products generated, go back to *E:\verification\EVS_output\New\GEFSv12_ens_mean_SQIN\GAXV2*, where all PNG and XML files will appear. The largest two files, ending with *pairs_raw* and *pairs_cond*, are generated from the run made on the *Verification* panel. All other files are generated from the run made on the *Output* panel. Verify that the RMSE plot,

GAXV2.GEFSv12_ens_mean_fcst_flow.Root_mean_square_error.png, is identical to Figure 60. In the figure, the *All data* curve is the same as the total uncertainty curve in Figure 35 in Chapter 7 except that the latter also shows the RMSE at lead time of 0, (that is, the analysis error). The user may now repeat the two remaining VUs on subpanel *1. Add verification unit(s)* to generate the other two RMSE results in Figure 35.

Many options and functions are available in each subpanel in the *Verification* panel under the *More* button (the user may have to scroll down to see the button). The user is encouraged to explore different options and functionalities available in the EVS using the above example in consultation with the EVS Manual.

A.3 Example 2 – Verification of ensemble streamflow forecast with aggregation of multiple forecast points

In this example, the user performs a reduced version of Case 3 in Chapter 7 in which only six forecast points are used (versus 26 in Case 3). The three downstream locations are BVDN4 (Delaware River at Belvidere, New Jersey, United States), MTMP1 (Delaware River at Matamoras/Port Jervis, New Jersey, United States) and WNTP1 (Lehigh River at Walnutport, Pennsylvania, United States), and the three headwater locations are HWYP1 (Lackawaxen River at Hawley, Pennsylvania, United States), WALN6 (West Branch Delaware River at Walton, New York, United States) and WHTP1 (Lehigh River at White Haven, Pennsylvania, United States).

To view the XML files for the individual ensemble forecasts, navigate to *E:\verification\Delaware\NAEFS_SQIN*, where the six folders for the six forecast points are located, each containing about three years' worth of NAEFS-forced streamflow ensemble forecasts. Go to *BVDN4_2018-2020* to list the individual files. Using a text editor or a viewer, open any one of the files. Verify that the headers at the top and bottom of the file show *ensemblememberid* of 0 and 41, respectively, indicating that this is a 42-member ensemble forecast. Make a note of *locationid* of BVDN4DEL and *parameterid* of QINE in the header. In the above, "DEL" denotes the Delaware River and "E" in QINE signifies that the instantaneous streamflow is estimated (that is, model-generated) rather than observed. The two identifiers, *locationid* and *parameterid*, will be recalled when the project file is described below.

There are two project files for Example 2, one for the downstream locations, *E:\verification\EVS_project_files\Delaware\NAEFS_SQIN_Downstream*, and the other for the headwater locations, *E:\verification\EVS_project_files\Delaware\NAEFS_SQIN_Headwater*. Double-click on *NAEFS_SQIN_Downstream* to launch the EVS. In subpanel *1. Add verification unit(s)*, the user will find three VUs for the three downstream locations. As in Example 1, the user must edit the paths for *Forecast data source* and *Observed data source* in subpanel *2b. Set input data sources* and *Folder for output statistics* in subpanel *2d. Set path for outputs* to match the actual folder locations on the user's computer.

If the user is adopting the EVS project files used for these examples as templates for verification of the user's own forecast, it is very important that *locationid* and *parameterid* in the header of the user's XML files match *Forecast data location id* and *Forecast data variable id*, respectively, in the *Other Options* tabbed pane in the *Input data options* subpanel, which is accessed through the *More* button in subpanel *2b. Set input data sources* in the *Verification* panel. If either of the identifiers does not match, the EVS will throw an error which can be difficult to trace.

To run the VU *MTMP1.NAEFS_fcst_flow*, click on the *Next* button in the lower-right corner of the *Verification* panel to get to subpanel *3a. Select metrics to compute*. Currently, only sample size, mean CRPS, ROC and reliability diagram are checked to reflect the metrics shown in Case 3 in Chapter 8. The user is encouraged to check or uncheck any metrics listed on subpanel *3a. Select metrics to compute*. To select all or none, right-click anywhere on the subpanel and click on the *select all/none* pop-up button.

Recall in Case 3 in Chapter 7 that CR decomposition was used for mean CRPS. To verify that this decomposition is included in the current VU, click on the element *Mean continuous ranked probability score* or *Ensemble distribution* in subpanel *3a. Select metrics to compute*. Then, in subpanel *3c. Edit basic parameters of selected metric 'Mean continuous ranked probability score'*, scroll down and click on the *More* button. On the *Main options* tabbed pane, verify that *Calibration-Refinement (CR)* is chosen for the *Select score decomposition:* option. Unlike the BS, which also allows for likelihood–base rate decomposition, mean CRPS allows only for CR decomposition. Note also in subpanel *3c. Edit basic parameters of selected metric 'Mean continuous ranked probability* score' that the highest threshold specified is 0.90 (versus 0.95 in Case 3 in Chapter 7), due to the much smaller sample size available from only three locations. Once the verification metrics are selected, click on the *Run* button to generate the two paired files in *E:\verification\EVS_output\Delaware\NAEFS_SQIN\MTMP1*.

Repeat the above steps for the other two downstream locations, WNTP1 and BVDN4. Or, if the user is certain that all three VUs are set up correctly and consistent, click the *Run all* button next to the *Run* button in the lower-left corner of the *Verification* panel to generate the two paired files for each location in *E:\verification\EVS_output\Delaware\NAEFS_SQIN*****, where the asterisks denote the 5-character location identifier.

Once the above run is complete, click on the *Next* button for the *Aggregation* panel, which should list the three VUs for the three locations in subpanel *2b. Select verification units to include in aggregation*. As noted in Case 3 in Chapter 7, the pairs for the three locations are to be pooled together for aggregated verification. To confirm this, click on the *More* button in this subpanel, verify that *Pool pairs (will ignore weight parameters)* is checked, and click on *OK* to close the pane. The verification statistics from aggregation are written into a separate folder, *E:\Verification\EVS_output\Delaware\NAEFS_SQIN\Downstream*. Verify in subpanel *2c. Set path for outputs* that *Folder for aggregated statistics* shows the above path. Click on the *Run* button near the lower-left corner. Once the run is complete, click on the *Next* button for the *Output* panel.

On the *Output* panel, subpanel *1a. Select unit(s) with results* should now list the three VUs for the individual forecast points followed by the aggregation unit *Downstream*. In this subpanel, click on *Downstream* or *Aggregation*, then right-click anywhere in subpanel *1b. Choose products for selected unit* and choose *Select all times and products*. Left-clicking on any of the product names will show all lead times in subpanel *1c. Choose lead times for selected product*. Click on the *Run* button to generate the products.

Once the above run is complete, navigate to *E:\verification\EVS_output\Delaware\NAEFS_SQIN\Downstream* to view all XML and PNG products and verify that *Downstream.Mean_continous_ranked_probability_score_SCORE.png* is identical to Figure 61. If desired, the EVS products for individual forecast points may also be generated. If generated, the location-specific products may be found in *E:\verification\EVS_output\Delaware\NAEFS_SQIN*****, where the asterisks denote the 5-character location identifier. Note that the sample size for the forecast point-specific results should be about one-third of that of the aggregation results. The user may repeat the above process for the headwater locations and verify that *Headwater.Mean_continous_ranked_probability_score_SCORE.png* in *E:\verification\EVS_output\Delaware\NAEFS_SQIN\Headwater* is identical to Figure 62.

The user is encouraged to explore other options and functionalities of the EVS for ensemble forecast verification in consultation with the EVS Manual. Depending on the options chosen, changes made to any of the variable or file names, and the number of runs made, the size of the output folders may become very large. It is hence a good practice to clean up the folders regularly.

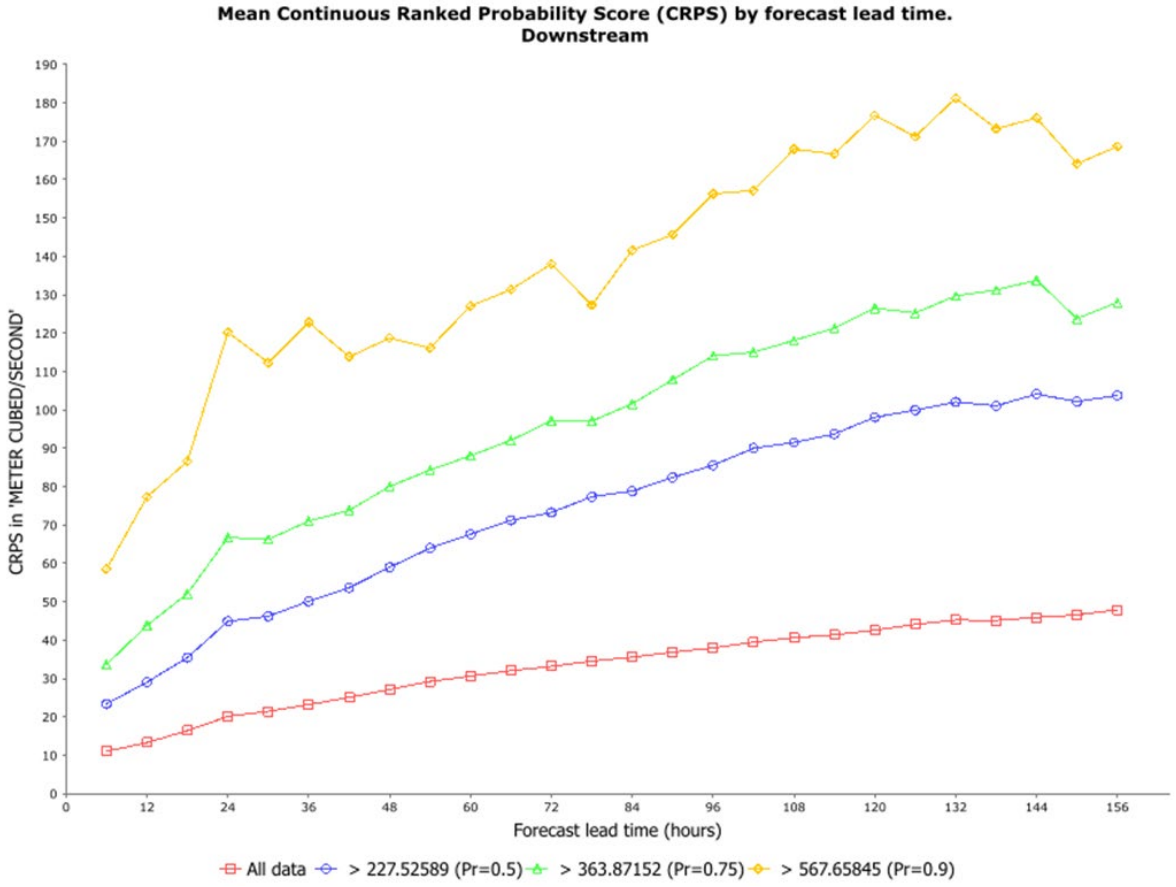


Figure 61. Mean CRPS result for downstream locations in Example 2

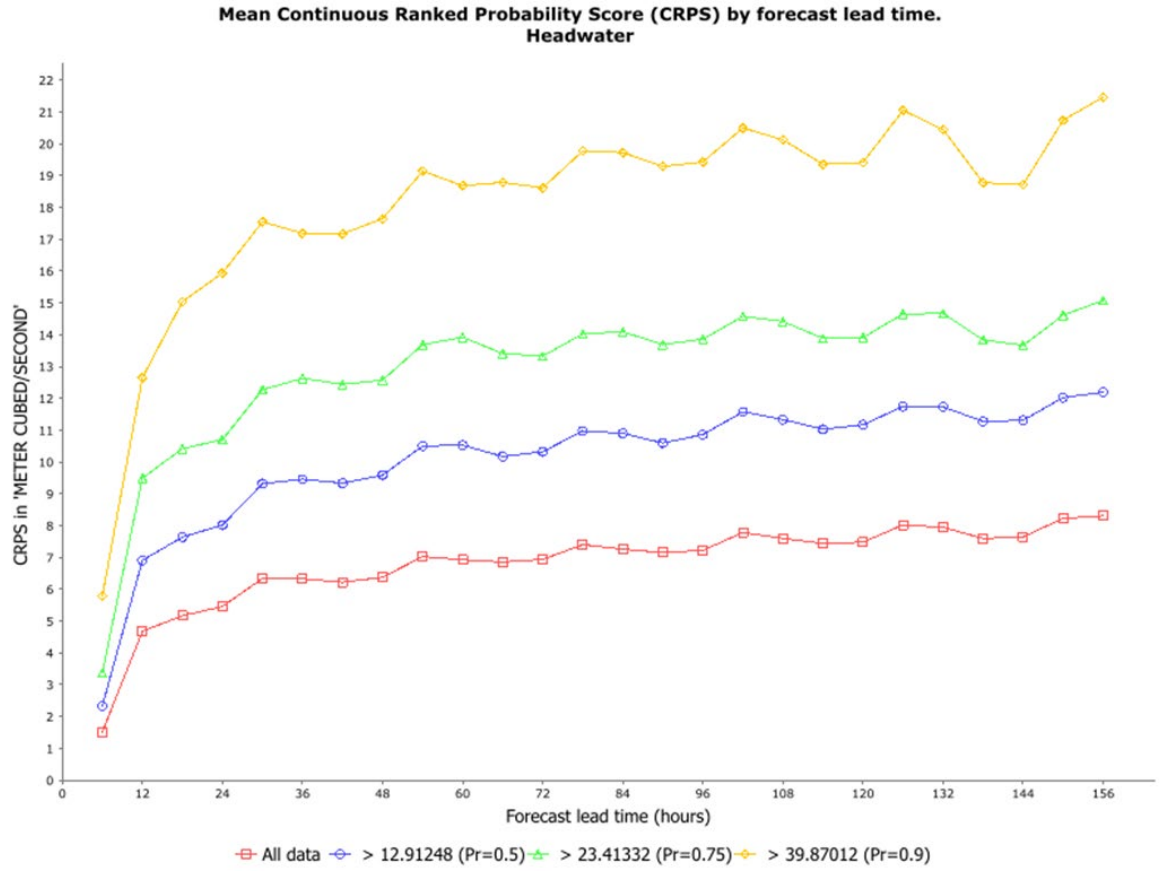


Figure 62. Mean CRPS result for headwater locations in Example 2

A.4 Example 3 – Verification of ensemble streamflow forecast with skill score and confidence interval calculations

In this example, the user performs a scaled-down version of Case 4 in Chapter 7 for verification of ensemble forecasts of daily-only streamflow for a single forecast point at SGET2 (Clear Creek near Sanger, Texas, United States) using a period of record of only one year (versus 31 years in Case 4). Due to the greatly reduced sample size, Example 3 is meant mainly to describe the mechanics of calculating skill scores and confidence intervals using the EVS, rather than to assess skill.

Navigate to *E:\verification\EVS_project_files\UTrinity* and double-click on *UTrinity_HEFS_SQIN*, which will launch the EVS for Example 3. Subpanel *1. Add verification unit(s)* should list three VUs. The VUs, *SGET2.MEFP_fcst_flow*, *SGET2.RClim_fcst_flow* and *SGET2.MEFP_EnsPost_fcst_flow*, are for the verification of the MEFP-forced ensemble streamflow forecast, resampled climatology and MEFP-forced ensemble streamflow forecasts with postprocessing (see Case 4 in Chapter 7 for explanation), respectively.

As in Examples 1 and 2, modify the paths to *Forecast data source*, *Observed data source* and *Folder for output statistics* to match their actual locations on the user's computer. In Example 3, *Observed data source* is also in XML. The user may want to verify that *Observed data location ID* and *Observed data variable ID* in the tabbed pane *Other options* (accessible through the *More* button in subpanel *2b. Set input data sources*) match *locationId* and *parameterId*, respectively, in the header of *E:\verification\UTrinity\QME\SGET2_QME.xml*. In the above, *QME* is NWS parlance for mean daily flow. Click on the *Next* button for the *Verification metrics to compute* panel.

On subpanel *3a. Select metrics to compute*, verify that only *Sample size* and *Mean continuous ranked probability skill score* are checked so that the user may assess the computing time for CRPSS calculation with confidence interval. If desired, the user may add any other metrics while on this subpanel.

On subpanel *3c. Edit basic parameters of selected metric 'Mean continuous ranked probability skill score'*, click on the *More* button. Verify in the *Main options* tabbed pane that *Reference forecast for skill* is set to *SGET2.RClim_fcst_flow* and click on the *Confidence intervals* tab. Verify that the technique is *stationary block bootstrap* and *Sample size*, *Minimum sample size*, *Average block size*, and *Units for block size* are *100*, *50*, *30.0* and *DAY*, respectively. For *Interval specification*, verify that *Lower and Upper* are set at *0.05* and *0.95*, respectively, for a 90% confidence interval. For explanations of the above options, the user is referred to pages 56 through 57 of the EVS Manual. Click on the *Run* button in the lower-left corner to generate the paired files.

Due to the confidence interval calculation, the above run takes longer than any of the runs in Examples 1 or 2. Once the run is complete, skip aggregation and go to the *Output* panel. At this point, the user may want to verify that the paired files have been generated in *E:\verification\EVS_output\UTrinity\MEFP_SQIN*, and that the file for raw pairs for the reference forecast, *RClim_SQIN*, has also been generated in *E:\verification\EVS_output\UTrinity\RClim_SQIN* for skill score calculation. Select *SGET2.MEFP_fcst_flow* in subpanel *1a. Select unit(s) with results* and right-click on any one of the products and choose *Select all times and products*. Left-clicking on any one of the products will list all applicable lead times in subpanel *1c. Choose lead times for selected product*. Click on the *Run* button in the lower-left corner to generate the products.

Navigate to *E:\verification\EVS_output\UTrinity\MEFP_SQIN* for the EVS products generated from the above run. Verify that *SGET2.MEFP_fcst_flow.Mean_continuous_ranked_probability_skill_score.png* is similar to Figure 63 (due to the randomness associated with bootstrapping, the result will not be identical). The figure shows that the CRPSS of MEFP-forced ensemble streamflow forecasts in reference to resampled climatology is positive at all lead times, but that, not surprisingly, the confidence intervals are very large due to the extremely short period of record. Repeat the above with the VU *SGET2.MEFP_EnsPost_fcst_flow* and navigate to *E:\verification\EVS_output\UTrinity\MEFP_EnsPost_SQIN* to verify that *SGET2.MEFP_EnsPost_fcst_flow.Mean_continuous_ranked_probability_skill_score.png* is similar to Figure 64. The figure shows that the CRPSS of MEFP-forced ensemble streamflow forecasts with EnsPost is larger than that without EnsPost (note that the y-axes are not identical between Figure 63 and Figure 64) at all lead times, but that the 90% confidence intervals are very large due to the very short period of record.

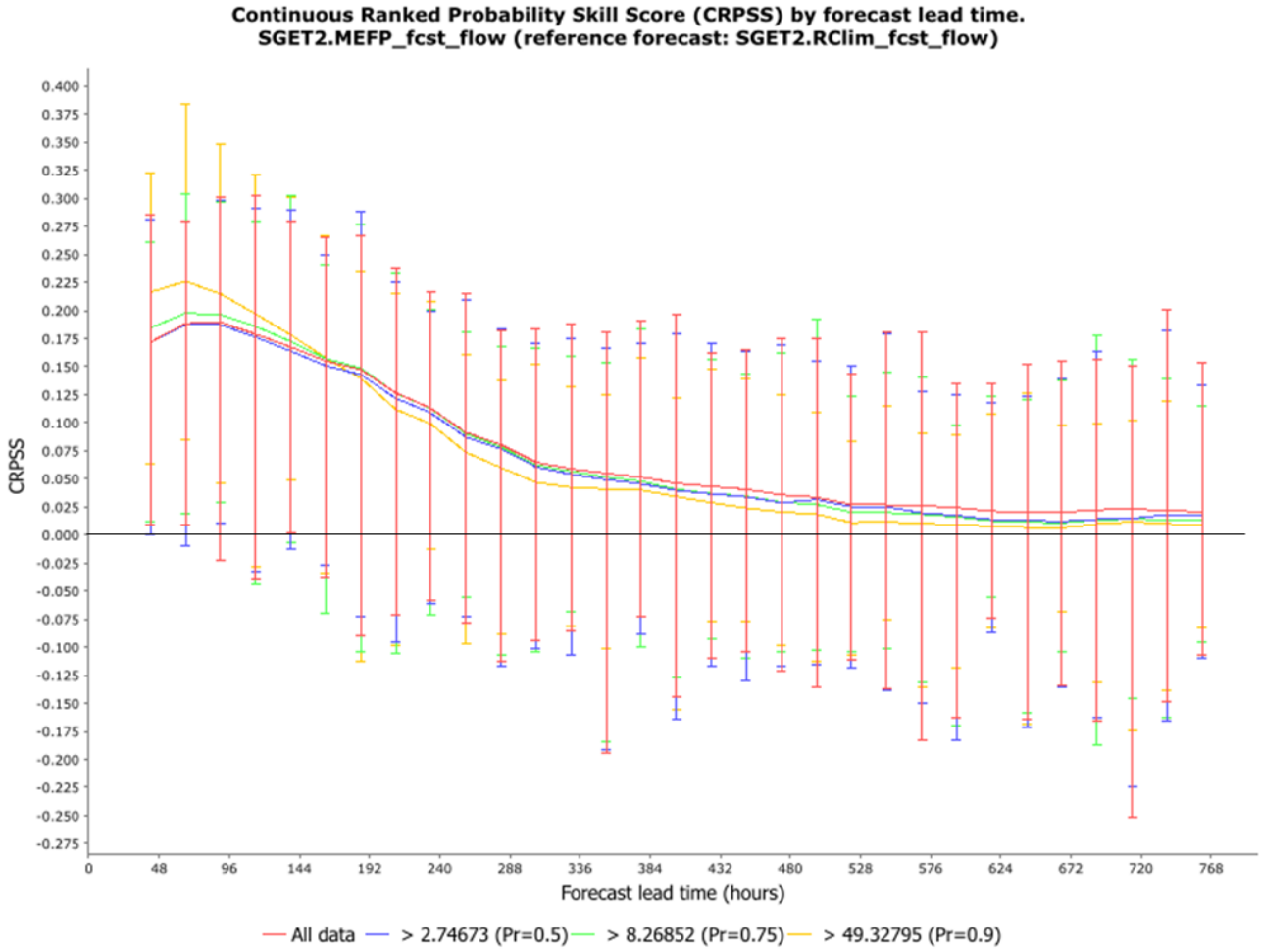


Figure 63. CRPSS with confidence intervals for MEFP-forced streamflow ensemble forecast for SGET2 in Example 3. The reference forecast is resampled climatology.

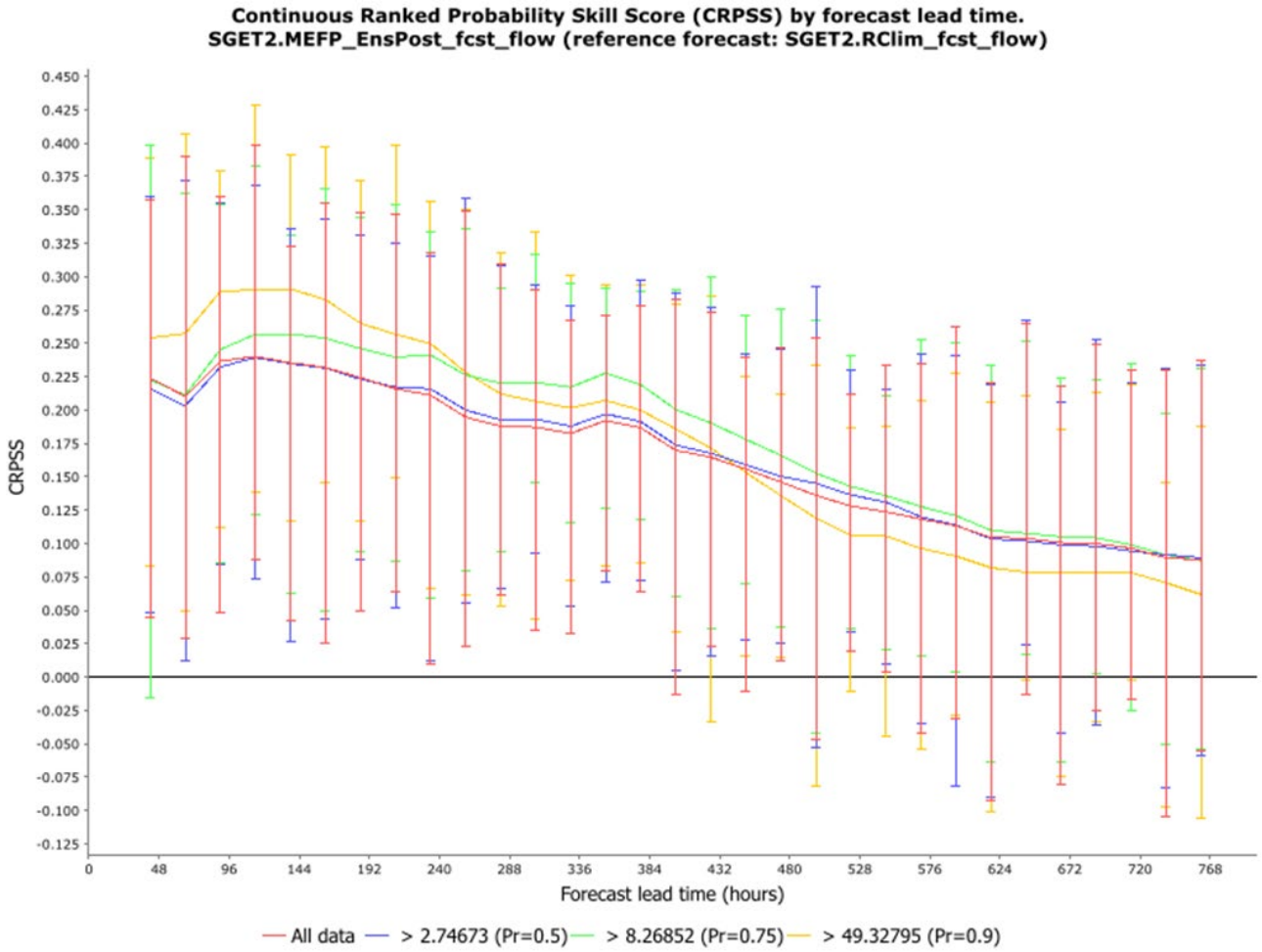


Figure 64. Same as Figure 63, but streamflow forecast is postprocessed with EnsPost

A.5 Examples 4–6: computational examples

Examples 4 through 6 use three different software tools to compute verification information for a single case study: ensemble streamflow forecasts for the Meuse River at St Pieter, just downstream of the Belgian–Dutch border. The examples are cast as "notebooks": documents from which code can be executed directly. For the printed version of the report, the HTML documents are included as appendices. Alongside the downloadable version of the present report on the WMO website, the actual notebooks as well as the data required to run the examples can be downloaded from the following link: https://wmostorage.blob.core.windows.net/wmo-public/hydro-forecast-verification-guidelines/A4-6_Computational_examples/A4-6_Computational_examples.zip.

A.6 Example 4 – Computational example: the Ensemble Verification System (EVS)

- The Ensemble Verification System
- EVS project file
- Data
- Running the EVS project
- Verification of deterministic forecasts
  - Mean absolute error
  - Mean error
  - Root mean square error

The Ensemble Verification System

The Ensemble Verification System (EVS) is designed to verify ensemble forecasts of hydrologic and hydrometeorological variables, such as temperature, precipitation, and streamflow, issued at discrete forecast locations: points or areas. The EVS can be downloaded from https://sourceforge.net/projects/ensemble-verification-system/. The download includes the EVS binaries, documentation and a sample project. The EVS requires a recent version of Java installed on the workstation where it is run. This makes the software platform independent: it can be run on Linux, MacOS as well as MS Windows.

The EVS can be run interactively (i.e., with a graphical user interface) as well as in command line mode. In the present example, the latter option will be used.

EVS project file

When running the EVS through its graphical user interface, it will save project settings to a `.evs` file which is xml-formatted. When running the EVS in command line mode, reference needs to be made to such a project file. For the purpose of the present example, a sample .evs file has been prepared.

```
<?xml version="1.0" standalone="yes"?><verification>
    <verification_unit>
        <identifiers>
            <location_id>H-MS-SINT</location_id>
            <environmental_variable_id>streamflow</environmental_variable_id>
            <additional_id>cosmo-leps</additional_id>
        </identifiers>
        <input_data>
            <forecast_data_source>
                <data>/home/jan/projecten/assembla-svn/wmo_computational_examples/evs/H-MS-SINT.fcst</data>
            </forecast_data_source>
            <observed_data_source>
                <data>/home/jan/projecten/assembla-svn/wmo_computational_examples/evs/H-MS-SINT.obs</data>
            </observed_data_source>
            <forecast_data_type>ASCII</forecast_data_type>
            <observed_data_type>ASCII</observed_data_type>
            <forecast_data_location_id>H-MS-SINT</forecast_data_location_id>
            <observed_data_location_id>H-MS-SINT</observed_data_location_id>
            <forecast_time_system>Coordinated Universal Time (UTC)</forecast_time_system>
            <observed_time_system>Coordinated Universal Time (UTC)</observed_time_system>
            <forecast_support>
```

```
                <statistic>INSTANTANEOUS</statistic>
                <existing_attribute_units>METER CUBED/SECOND</existing_attribute_units>
                <notes></notes>
            </forecast_support>
            <observed_support>
                <statistic>INSTANTANEOUS</statistic>
                <existing_attribute_units>METER CUBED/SECOND</existing_attribute_units>
                <notes></notes>
            </observed_support>
            <use_all_observations_for_climatology>false</use_all_observations_for_climato
logy>
            <apply_date_cond_to_climatology>false</apply_date_cond_to_climatology>
            <apply_value_cond_to_climatology>false</apply_value_cond_to_climatology>
            <forecast_date_format>yyyyMMddHHmm</forecast_date_format>
            <observed_date_format>yyyyMMddHHmm</observed_date_format>
            <global_null_value>-999.0</global_null_value>
            <data_services>
                <data_service>
                    <name>FEWS-DS</name>
                    <forecast_filterId>simulated</forecast_filterId>
                    <observed_filterId>observed</observed_filterId>
                    <convertDatum>false</convertDatum>
                    <useDisplayUnits>false</useDisplayUnits>
                    <showThresholds>false</showThresholds>
                    <omitMissing>true</omitMissing>
                    <onlyHeaders>false</onlyHeaders>
                    <documentVersion>1.23</documentVersion>
                    <forecastCount>1000000</forecastCount>
                </data_service>
            </data_services>
        </input_data>
        <verification_window>
            <start_date>
                <year>1900</year>
                <month>0</month>
                <day>1</day>
            </start_date>
            <end_date>
                <year>2100</year>
                <month>0</month>
                <day>1</day>
```

```
            </end_date>
            <window_in_valid_time>true</window_in_valid_time>
            <first_lead_period>48.0</first_lead_period>
            <last_lead_period>48.0</last_lead_period>
            <forecast_lead_units>HOUR</forecast_lead_units>
            <sample_size_constraint>0.0</sample_size_constraint>
        </verification_window>
        <output_data>
            <output_data_location>.</output_data_location>
            <output_graphics_type>PNG</output_graphics_type>
        </output_data>
        <paired_data>
            <paired_data_location>/home/jan/projecten/assembla-svn/wmo_computational_exam
ples/evs/H-MS-SINT_streamflow_cosmo-leps_pairs_raw.xml</paired_data_location>
            <eliminate_duplicates>true</eliminate_duplicates>
            <write_conditional_pairs>true</write_conditional_pairs>
            <write_unconditional_pairs>true</write_unconditional_pairs>
            <write_gzip_pairs>false</write_gzip_pairs>
            <paired_write_precision>5</paired_write_precision>
            <strip_nulls_from_paired_file>true</strip_nulls_from_paired_file>
            <raw_pairs_in_aggregated_res>false</raw_pairs_in_aggregated_res>
        </paired_data>
        <metrics>
            <metric>
                <name>BrierScore</name>
                <double_array_parameter>200.0</double_array_parameter>
                <main_threshold>true</main_threshold>
                <threshold_condition>isGreater</threshold_condition>
                <decompose_parameter>NONE</decompose_parameter>
                <forecast_type_parameter>regular</forecast_type_parameter>
                <unconditional_parameter>false</unconditional_parameter>
                <minimum_sample_size_parameter>0</minimum_sample_size_parameter>
                <bootstrap_parameters>
                    <technique>None</technique>
                </bootstrap_parameters>
            </metric>
            <metric>
                <name>MeanAbsoluteError</name>
                <probability_array_parameter>-Infinity</probability_array_parameter>
                <main_threshold>true</main_threshold>
                <threshold_condition>isGreater</threshold_condition>
```

```
            <forecast_type_parameter>regular</forecast_type_parameter>
            <unconditional_parameter>false</unconditional_parameter>
            <forecast_average_parameter>Mean</forecast_average_parameter>
            <minimum_sample_size_parameter>0</minimum_sample_size_parameter>
            <bootstrap_parameters>
                <technique>None</technique>
            </bootstrap_parameters>
        </metric>
        <metric>
            <name>MeanError</name>
            <double_array_parameter>-Infinity</double_array_parameter>
            <main_threshold>true</main_threshold>
            <threshold_condition>isGreater</threshold_condition>
            <forecast_type_parameter>regular</forecast_type_parameter>
            <unconditional_parameter>false</unconditional_parameter>
            <forecast_average_parameter>Mean</forecast_average_parameter>
            <minimum_sample_size_parameter>0</minimum_sample_size_parameter>
            <bootstrap_parameters>
                <technique>None</technique>
            </bootstrap_parameters>
        </metric>
        <metric>
            <name>RelativeOperatingCharacteristic</name>
            <double_array_parameter>200.0</double_array_parameter>
            <main_threshold>true</main_threshold>
            <threshold_condition>isGreater</threshold_condition>
            <roc_points_parameter>10</roc_points_parameter>
            <forecast_type_parameter>regular</forecast_type_parameter>
            <unconditional_parameter>false</unconditional_parameter>
            <fitted_roc_parameter>false</fitted_roc_parameter>
            <minimum_sample_size_parameter>0</minimum_sample_size_parameter>
            <bootstrap_parameters>
                <technique>None</technique>
            </bootstrap_parameters>
        </metric>
        <metric>
            <name>ReliabilityDiagram</name>
            <double_array_parameter>200.0</double_array_parameter>
            <main_threshold>true</main_threshold>
            <threshold_condition>isGreater</threshold_condition>
```

```
                <forecast_type_parameter>regular</forecast_type_parameter>
                <unconditional_parameter>false</unconditional_parameter>
                <equal_samples_parameter>false</equal_samples_parameter>
                <reliability_points_parameter>5</reliability_points_parameter>
                <minimum_sample_size_parameter>0</minimum_sample_size_parameter>
                <bootstrap_parameters>
                    <technique>None</technique>
                </bootstrap_parameters>
            </metric>
            <metric>
                <name>RootMeanSquareError</name>
                <double_array_parameter>-Infinity</double_array_parameter>
                <main_threshold>true</main_threshold>
                <threshold_condition>isGreater</threshold_condition>
                <forecast_type_parameter>regular</forecast_type_parameter>
                <unconditional_parameter>false</unconditional_parameter>
                <forecast_average_parameter>Mean</forecast_average_parameter>
                <minimum_sample_size_parameter>0</minimum_sample_size_parameter>
                <bootstrap_parameters>
                    <technique>None</technique>
                </bootstrap_parameters>
            </metric>
            <metric>
                <name>SampleSize</name>
                <double_array_parameter>-Infinity</double_array_parameter>
                <main_threshold>true</main_threshold>
                <threshold_condition>isGreater</threshold_condition>
                <forecast_type_parameter>regular</forecast_type_parameter>
                <unconditional_parameter>false</unconditional_parameter>
            </metric>
        </metrics>
    </verification_unit>
</verification>
```

The example project file contains information for the EVS to compute various measures and metrics. Note that, for the purpose of the present example, these are computed at the 48-hour lead time only. Often, one will want to compute verification data for more than a single lead time.

Data

The data that is referred to in the EVS project file comprises two files: a file containing forecast data (`H-MS-SINT.fcst`) and a file containing observed data (`H-MS-SINT.obs`).

The file containing observations contains two columns: one for the valid time and one for forecast lead time.

```
cat(readLines('H-MS-SINT.obs',n=10),sep='\n')
```

```
200801011200 238.12

200801011500 236.11

200801011800 244 .53

200801012100 242.59

200801020000 245.33

200801020600 249.64

200801021200 230.03

200801021800 172.21

200801030000 159.24

200801030600 247.9
```

The forecasts are sixteen-member ensemble forecasts. The file contains 18 columns: forecast valid time, forecast lead time and 16 columns, each containing a forecast from a single ensemble member.

```
cat(readLines('H-MS-SINT.fcst',n=10),sep='\n')
```

```
200801011200 0 238.12 238.12 238.12 238.12 238.12 238.12 238.12 238.12 238.12 238.12 238.
12 238.12 238.12 238.12 238.12 238.12

200801011500 3 235.98 235.98 235.98 235.98 235.98 235.98 235.98 235.98 235.98 235.99 235.
98 235.98 235.98 235.98 235.98 235.98

200801011800 6 216.95 216.95 216.95 216.95 216.95 216.95 217.08 216.95 216.95 216.96 216.
95 216.95 216.95 216.95 216.97 216.95

200801012100 9 220.4 220.4 220.4 220.4 220.4 220.4 221.03 220.4 220.41 220.45 220.4 220.3
9 220.4 220.4 220.5 220.4

200801020000 12 203.19 203.08 203.08 203.18 203.21 203.18 204.05 203.2 203.18 203.19 203.
08 203.15 203.17 203.21 203.36 203.18

200801020600 18 212.52 211.86 211.86 212.43 212.69 212.38 213.95 212.47 212.35 212.42 211
.86 212.25 212.39 212.71 213.14 212.4

200801021200 24 212.37 211.87 211.87 212.28 212.57 212.25 213.85 212.31 212.21 212.48 211
.87 212.14 212.28 212.58 213 212.21

200801021800 30 202.2 201.7 201.7 202.11 202.37 202.11 203.57 202.14 202.05 202.42 201.7
201.97 202.13 202.39 202.77 202.07

200801030000 36 182.67 182.23 182.22 182.59 182.82 182.63 183.93 182.61 182.54 182.9 182.
23 182.47 182.61 182.85 183.19 182.56

200801030600 42 196.9 196.47 196.47 196.82 197.05 196.89 198.15 196.85 196.77 197.07 196.
47 196.7 196.85 197.08 197.42 196.79
```

Running the EVS project

The EVS project file is run by referring to the workstation's Java and to the project file.

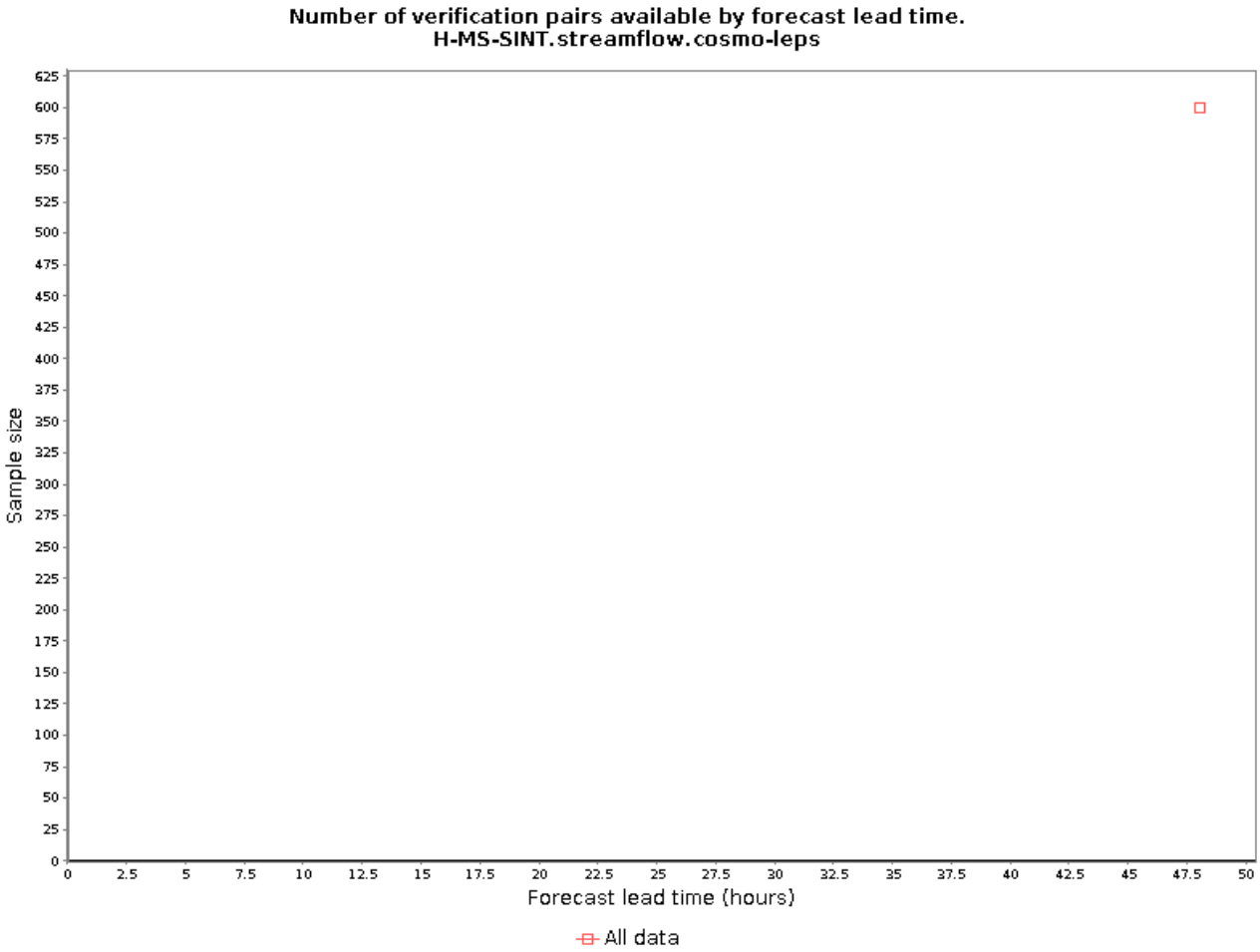
```
java -jar EVS.jar computational_example_evs.evs
```

The run will have produced various files including graphs that show verification results and xml files containing numerical values. These include files that give information about the sample size: an xml file and an image file.

```
<?xml version="1.0" standalone="yes"?><results>
<!--
    Result file containing the results for a single metric by lead period.
    Some metrics, such as reliability diagrams, have results for specific thresholds
    (e.g. probability thresholds).  In that case, the results are stored by lead period
    and then by threshold value.  The actual data associated with a result always appears
    within a 'values' tag.  A metric result that comprises a single value will appear as
    a single value in this tag.  A metric result that comprises a 1D matrix will appear
    as a row of values separated by commas in the input order.  A metric result that
    comprises a 2D matrix will appear as a sequence of rows, each with a 'values' tag,
    which are written in the input order.  For example, a diagram metric with an x and y
    axis will comprise two rows of data (i.e. two rows within two separate 'values'
    tags).  The default input order would be data for the x axis followed by data for the
    y axis.  Data that refer to cumulative probabilities are, by default, always defined
    in increasing size of probability.
-->
    <meta_data>
        <thresholds_type>true</thresholds_type>
        <original_file_id>H-MS-SINT.streamflow.cosmo-leps.Sample_size.xml</original_file_
id>
     </meta_data>
    <result>
        <lead_hour>48.0</lead_hour>
        <threshold_data>
            <threshold>
                <threshold_value>All data</threshold_value>
                <data>
                    <values>600.0</values>
                </data>
            </threshold>
        </threshold_data>
    </result>
</results>
```

In graphical format, the information looks as follows.

```
echo "![](H-MS-SINT.streamflow.cosmo-leps.Sample_size.png)";
```



Verification of deterministic forecasts

The deterministic verification metrics that are computed, are:

- Mean absolute error
- Mean error
- Root mean square error

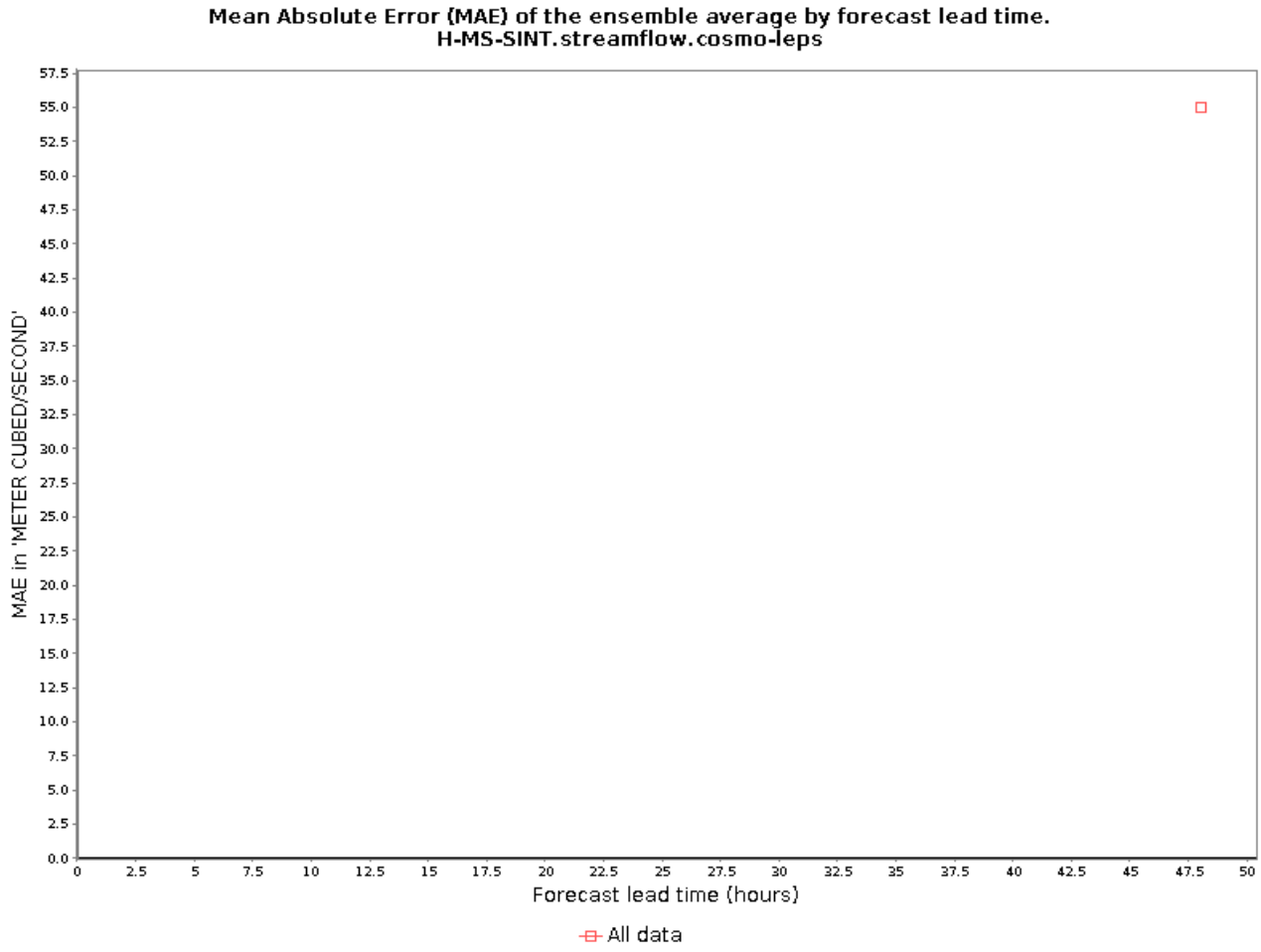
The EVS project is set up to, where deterministic forecasts are verified, use the ensemble mean as the deterministic forecast.

Mean absolute error

From the xml and the png files, we identify the value of the mean absolute error as approx. 55 m33/s.

```
<?xml version="1.0" standalone="yes"?><results>
<!--
    Result file containing the results for a single metric by lead period.
    Some metrics, such as reliability diagrams, have results for specific thresholds
    (e.g. probability thresholds).  In that case, the results are stored by lead period
    and then by threshold value.  The actual data associated with a result always appears
    within a 'values' tag.  A metric result that comprises a single value will appear as
    a single value in this tag.  A metric result that comprises a 1D matrix will appear
    as a row of values separated by commas in the input order.  A metric result that
```

```
    comprises a 2D matrix will appear as a sequence of rows, each with a 'values' tag,
    which are written in the input order.  For example, a diagram metric with an x and y
    axis will comprise two rows of data (i.e. two rows within two separate 'values'
    tags).  The default input order would be data for the x axis followed by data for the
    y axis.  Data that refer to cumulative probabilities are, by default, always defined
    in increasing size of probability.
-->
    <meta_data>
        <thresholds_type>true</thresholds_type>
        <original_file_id>H-MS-SINT.streamflow.cosmo-leps.Mean_absolute_error.xml</origin
al_file_id>
    </meta_data>
    <result>
        <lead_hour>48.0</lead_hour>
        <threshold_data>
            <threshold>
                <threshold_value>All data</threshold_value>
                  <data>
                      <values>55.004348958333345</values>
                  </data>
             </threshold>
         </threshold_data>
    </result>
</results>
echo "![](H-MS-SINT.streamflow.cosmo-leps.Mean_absolute_error.png)";
```

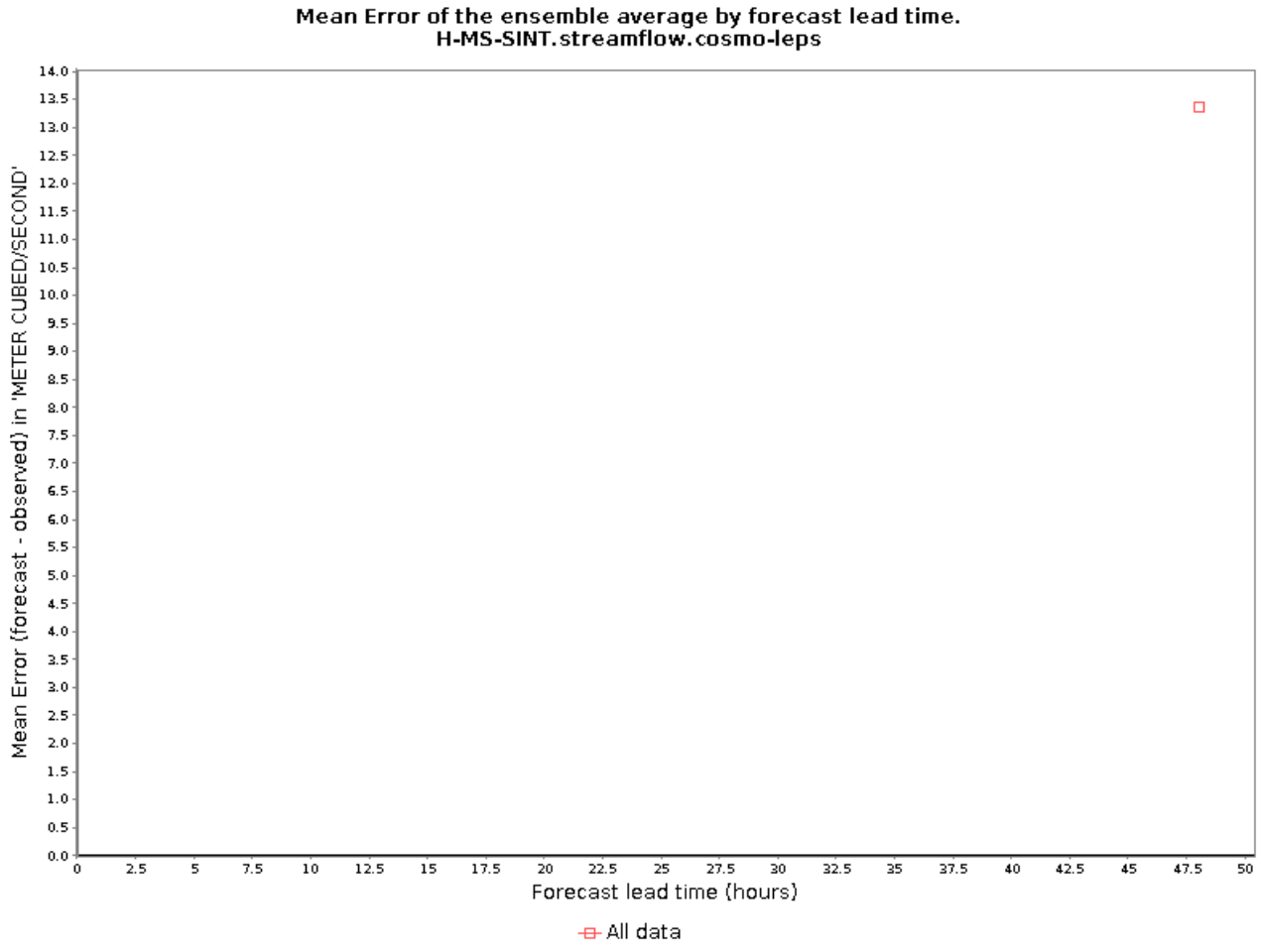


Mean error

From the xml and the png files, we identify the value of the mean error as 13.4 m3/s.

```
<?xml version="1.0" standalone="yes"?><results>
<!--
    Result file containing the results for a single metric by lead period.
    Some metrics, such as reliability diagrams, have results for specific thresholds
    (e.g. probability thresholds).  In that case, the results are stored by lead period
    and then by threshold value.  The actual data associated with a result always appears
    within a 'values' tag.  A metric result that comprises a single value will appear as
    a single value in this tag.  A metric result that comprises a 1D matrix will appear
    as a row of values separated by commas in the input order.  A metric result that
    comprises a 2D matrix will appear as a sequence of rows, each with a 'values' tag,
    which are written in the input order.  For example, a diagram metric with an x and y
    axis will comprise two rows of data (i.e. two rows within two separate 'values'
    tags).  The default input order would be data for the x axis followed by data for the
    y axis.  Data that refer to cumulative probabilities are, by default, always defined
    in increasing size of probability.
-->
    <meta_data>
```

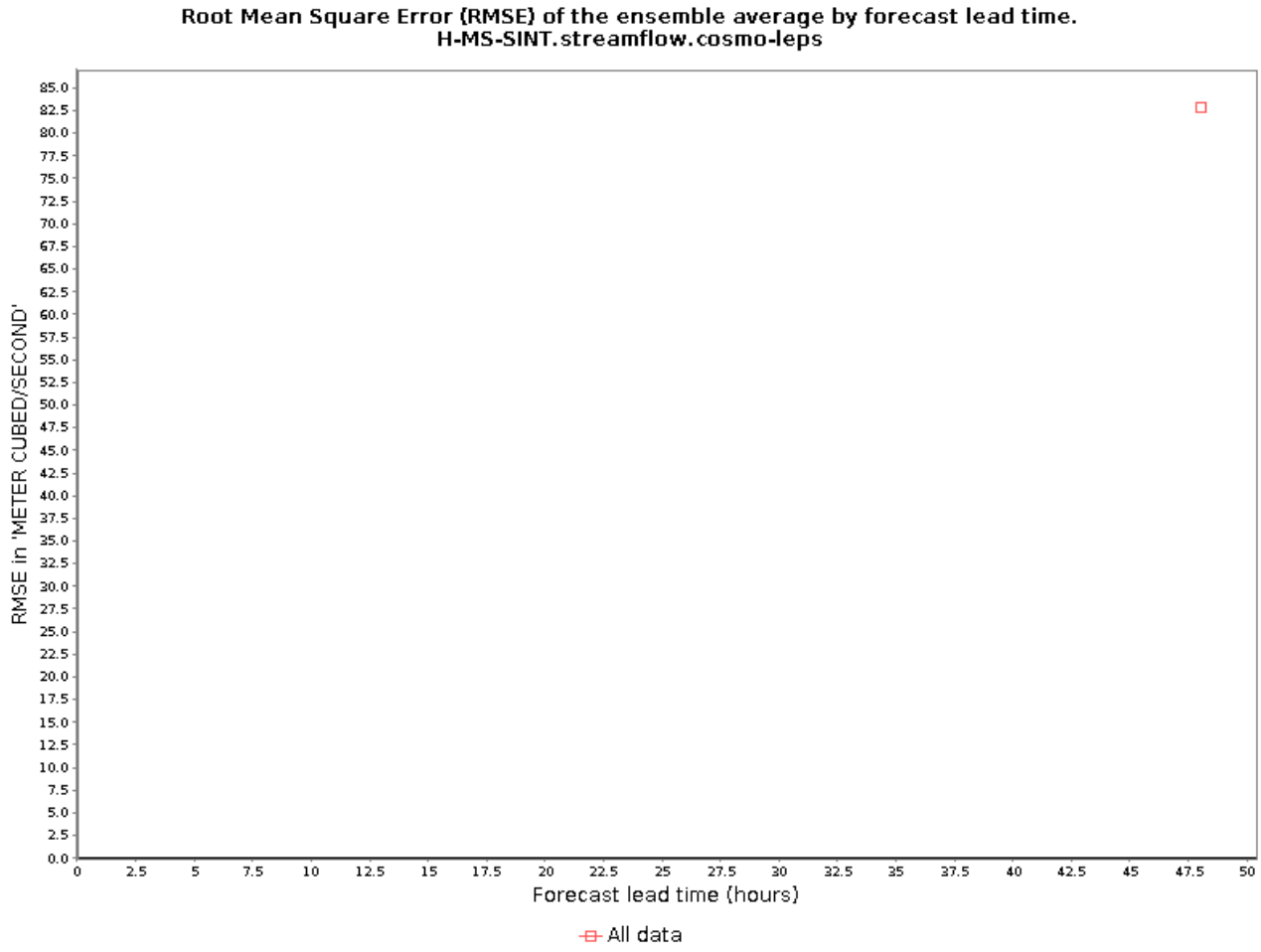
```
        <thresholds_type>true</thresholds_type>
        <original_file_id>H-MS-SINT.streamflow.cosmo-leps.Mean_error.xml</original_file_i
d>
    </meta_data>
    <result>
        <lead_hour>48.0</lead_hour>
        <threshold_data>
            <threshold>
                <threshold_value>All data</threshold_value>
                <data>
                    <values>13.366328125000004</values>
                </data>
            </threshold>
        </threshold_data>
    </result>
</results>
echo "![](H-MS-SINT.streamflow.cosmo-leps.Mean_error.png)";
```



Root mean square error

From the xml ad the png files, we identify the value of the root mean square error as 82.9 m3/s.

```
<?xml version="1.0" standalone="yes"?><results>
<!--
    Result file containing the results for a single metric by lead period.
    Some metrics, such as reliability diagrams, have results for specific thresholds
    (e.g. probability thresholds).  In that case, the results are stored by lead period
    and then by threshold value.  The actual data associated with a result always appears
    within a 'values' tag.  A metric result that comprises a single value will appear as
    a single value in this tag.  A metric result that comprises a 1D matrix will appear
    as a row of values separated by commas in the input order.  A metric result that
    comprises a 2D matrix will appear as a sequence of rows, each with a 'values' tag,
    which are written in the input order.  For example, a diagram metric with an x and y
    axis will comprise two rows of data (i.e. two rows within two separate 'values'
    tags).  The default input order would be data for the x axis followed by data for the
    y axis.  Data that refer to cumulative probabilities are, by default, always defined
    in increasing size of probability.
-->
    <meta_data>
        <thresholds_type>true</thresholds_type>
        <original_file_id>H-MS-SINT.streamflow.cosmo-leps.Root_mean_square_error.xml</ori
ginal_file_id>
    </meta_data>
    <result>
        <lead_hour>48.0</lead_hour>
        <threshold_data>
            <threshold>
                <threshold_value>All data</threshold_value>
                <data>
                    <values>82.8658049499192</values>
                </data>
            </threshold>
        </threshold_data>
    </result>
</results>
echo "![](H-MS-SINT.streamflow.cosmo-leps.Root_mean_square_error.png)";
```



Verification of probabilistic forecasts

The probabilistic verification metrics that are computed, are:

- Brier's probability score
- Reliability diagram
- Relative Operating Characteristic

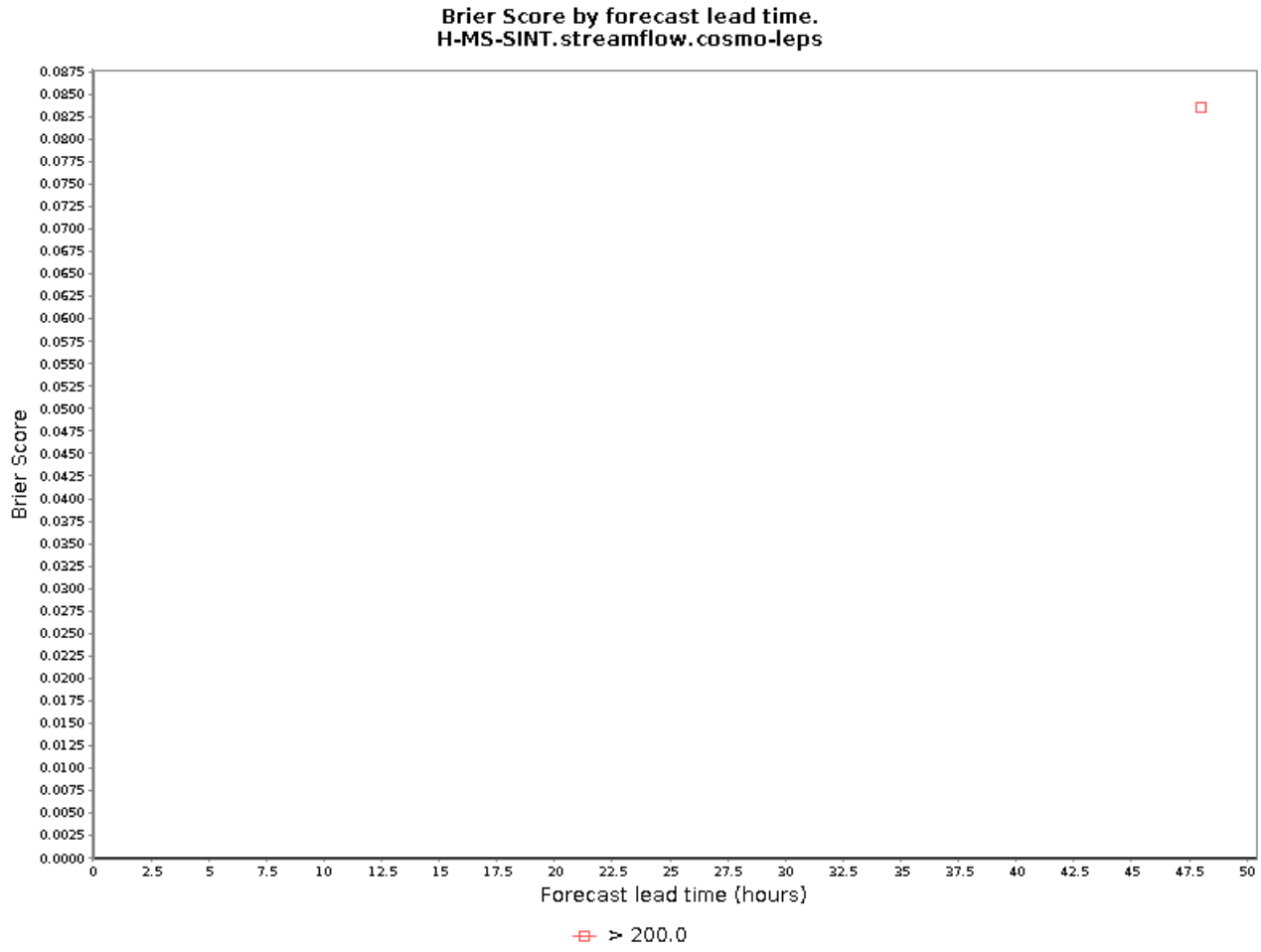
These metrics are computed for the event that streamflow, either forecasted or observed, exceeds the 200 m33/s threshold.

Brier's probability score

Brier's probability score is computed for the streamflow event defined by the exceedance of the 200 m33/s threshold. The (unitless) value of the score is 0.08.

```
<?xml version="1.0" standalone="yes"?><results>
<!--
    Result file containing the results for a single metric by lead period.
    Some metrics, such as reliability diagrams, have results for specific thresholds
    (e.g. probability thresholds).  In that case, the results are stored by lead period
    and then by threshold value.  The actual data associated with a result always appears
     within a 'values' tag.  A metric result that comprises a single value will
     appear as a single value in this tag.  A metric result that comprises a 1D
```

```
    matrix will appear as a row of values separated by commas in the input order.  A
    metric result that comprises a 2D matrix will appear as a sequence of rows, each with
    a 'values' tag, which are written in the input order.  For example, a diagram metric
    with an x and y axis will comprise two rows of data (i.e. two rows within two
    separate 'values' tags).  The default input order would be data for the x axis
    followed by data for the y axis.  Data that refer to cumulative probabilities are, by
    default, always defined in increasing size of probability.
-->
    <meta_data>
        <thresholds_type>true</thresholds_type>
        <original_file_id>H-MS-SINT.streamflow.cosmo-leps.Brier_score.xml</original_file_
id>
    </meta_data>
    <result>
        <lead_hour>48.0</lead_hour>
        <threshold_data>
            <threshold>
                <threshold_value>GT 200.0</threshold_value>
                <data>
                    <values>0.08354817708333333</values>
                </data>
            </threshold>
        </threshold_data>
    </result>
</results>
echo "![](H-MS-SINT.streamflow.cosmo-leps.Brier_score.png)";
```

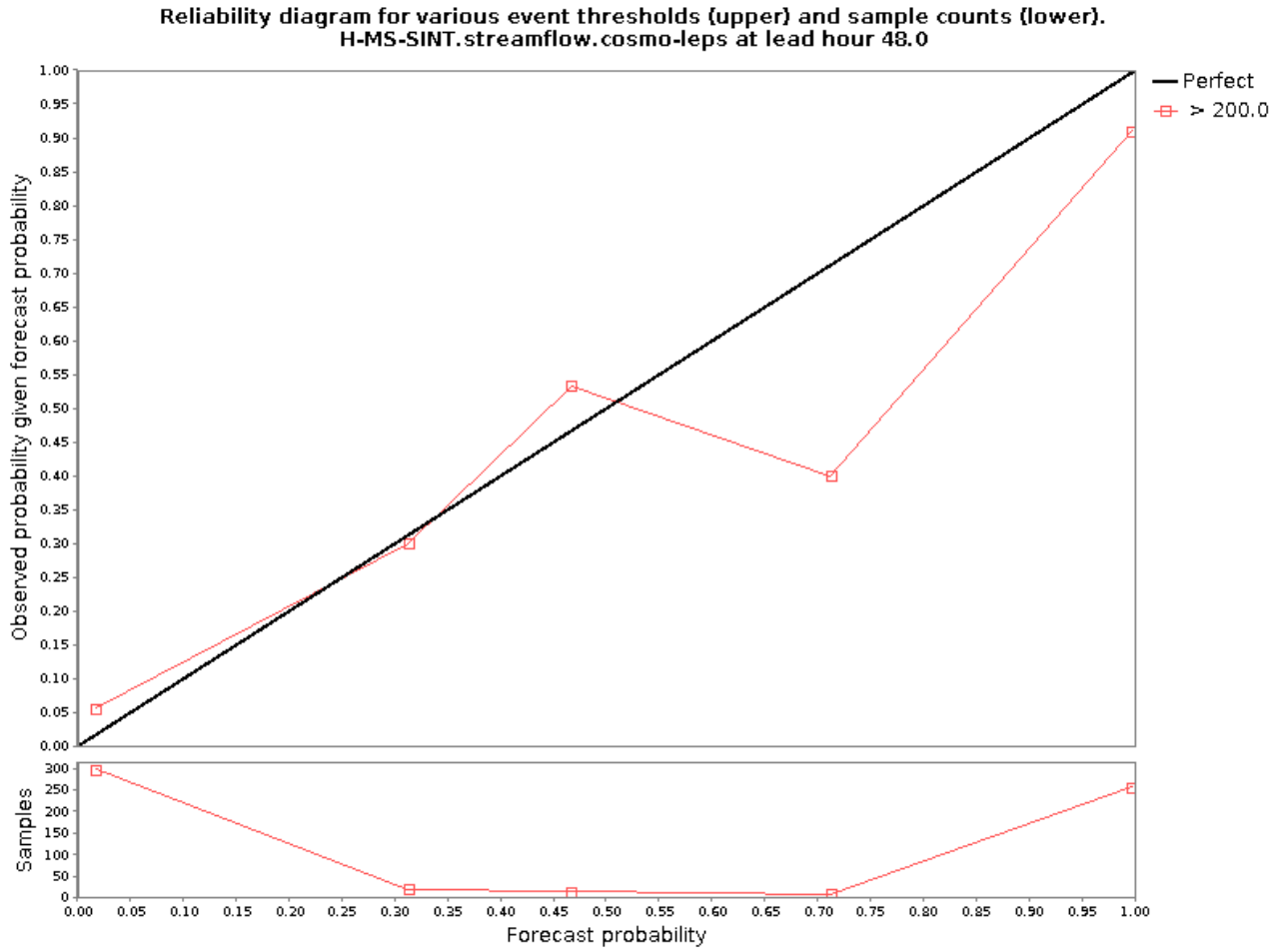


Reliability diagram

The reliability diagram consists of multiple data points (or plotting positions) which are included in the xml output.

```
<?xml version="1.0" standalone="yes"?><results>
<!--
    Result file containing the results for a single metric by lead period.
    Some metrics, such as reliability diagrams, have results for specific thresholds
    (e.g. probability thresholds).  In that case, the results are stored by lead period
    and then by threshold value.  The actual data associated with a result always appears
    within a 'values' tag.  A metric result that comprises a single value will appear as
    a single value in this tag.  A metric result that comprises a 1D matrix will appear
    as a row of values separated by commas in the input order.  A metric result that
    comprises a 2D matrix will appear as a sequence of rows, each with a 'values' tag,
    which are written in the input order.  For example, a diagram metric with an x and y
    axis will comprise two rows of data (i.e. two rows within two separate 'values'
    tags).  The default input order would be data for the x axis followed by data for the
    y axis.  Data that refer to cumulative probabilities are, by default, always defined
    in increasing size of probability.
-->
```

```
    <meta_data>
        <thresholds_type>true</thresholds_type>
        <original_file_id>H-MS-SINT.streamflow.cosmo-leps.Reliability_diagram.xml</origin
al_file_id>
    </meta_data>
    <result>
        <lead_hour>48.0</lead_hour>
        <threshold_data>
            <threshold>
                <threshold_value>GT 200.0</threshold_value>
                <data>
                    <values>0.01741, 0.3125, 0.46667, 0.7125, 0.99708</values>
                    <values>0.05705, 0.3, 0.53333, 0.4, 0.91051</values>
                    <values>298.0, 20.0, 15.0, 10.0, 257.0</values>
                    <values>0.44833, 0.44833, 0.44833, 0.44833, 0.44833</values>
                    <values>0.23287, 0.38042, 0.4575, 0.58042, 0.72271</values>
                </data>
            </threshold>
        </threshold_data>
    </result>
</results>
echo "![](H-MS-SINT.streamflow.cosmo-leps.Reliability_diagram.48.0.png)";
```

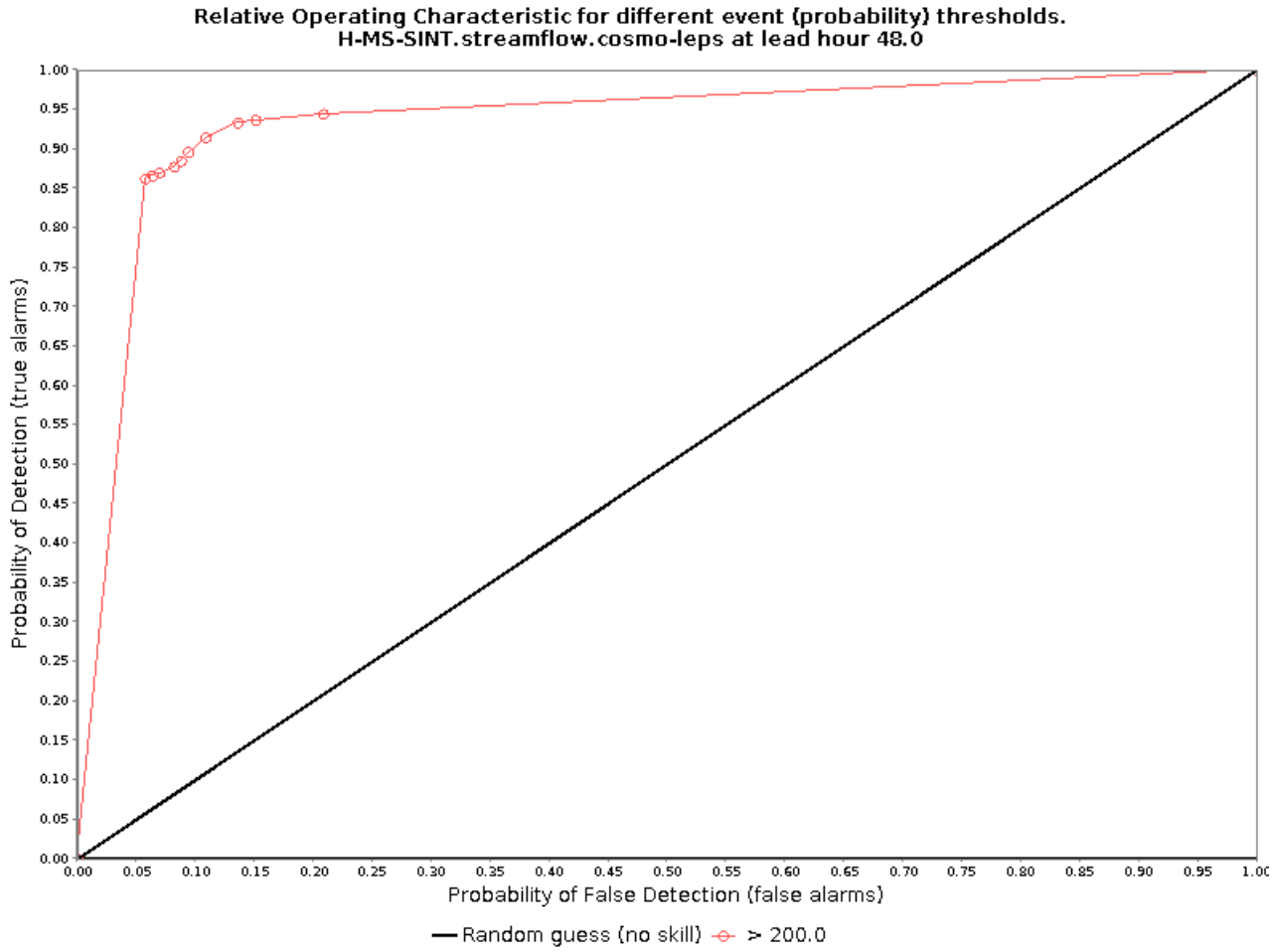


Relative Operating Characteristic

The relative operating characteristic, too, consists of multiple data points (or plotting positions) which are included in the xml output.

```
<?xml version="1.0" standalone="yes"?><results>
<!--
    Result file containing the results for a single metric by lead period.
    Some metrics, such as reliability diagrams, have results for specific thresholds
    (e.g. probability thresholds).  In that case, the results are stored by lead period
    and then by threshold value.  The actual data associated with a result always appears
    within a 'values' tag.  A metric result that comprises a single value will appear as
    a single value in this tag.  A metric result that comprises a 1D matrix will appear
    as a row of values separated by commas in the input order.  A metric result that
   comprises a 2D matrix will appear as a sequence of rows, each with a 'values' tag,
    which are written in the input order.  For example, a diagram metric with an x and y
    axis will comprise two rows of data (i.e. two rows within two separate 'values'
    tags).  The default input order would be data for the x axis followed by data for the
    y axis.  Data that refer to cumulative probabilities are, by default, always defined
    in increasing size of probability.
-->
```

```
    <meta_data>
        <thresholds_type>true</thresholds_type>
        <original_file_id>H-MS-SINT.streamflow.cosmo-leps.Relative_operating_characteristic.xml</original_file_id>
    </meta_data>
    <result>
        <lead_hour>48.0</lead_hour>
        <threshold_data>
            <threshold>
                <threshold_value>GT 200.0</threshold_value>
                <data>
                    <values>0.0, 0.0574, 0.06344, 0.06949, 0.08157, 0.08761, 0.09366, 0.10876, 0.13595, 0.15106, 0.20846, 1.0, 1.0</values>
                    <values>0.0, 0.86245, 0.86617, 0.86989, 0.87732, 0.88476, 0.89591, 0.9145, 0.93309, 0.9368, 0.94424, 1.0, 1.0</values>
                </data>
            </threshold>
        </threshold_data>
    </result>
</results>
echo "![](H-MS-SINT.streamflow.cosmo-leps.Relative_operating_characteristic.48.0.png)";
```



Parsing EVS outputs to R

The EVS xml outputs can be parsed to R using a script that is part of the EVS download: `Utilities.R`. The script requires the R `XML` package to be installed.

```
source('Utilities.R')
```

The script includes various functions including `readEVSScores` for reading single-valued verification metrics, and `readEVSDiagrams` for reading data for various technical diagrams. The functions store the data from the xml file into a list, from which it can be read for further use.

```
readEVSScores(file='H-MS-SINT.streamflow.cosmo-leps.Brier_score.xml')
## $lead.times
## [1] 48
##
## $events
## [1] "> 200.0"
##
## $events.numeric
## [1] 200
##
## $events.probs
```

```
## [1] NA
##
## $scores
##             [,1]
## [1,] 0.08354818
##
## $lower
##      [,1]
## [1,]   NA
##
## $upper
##      [,1]
## [1,]   NA
```

A.7 Example 5 – Computational example: the R verification package

The present document contains a computational verification example using R and its `verification` package. The example includes various data processing steps. These are done using the `reshape2`, `lubridate`, and `dplyr` packages. These processing steps will not be explained in detail. Plotting is done using the `ggplot2` package. You'll need to have these pre-installed. The file includes code for creating an animation. That code is not run. If you do want to run it, you'll have to ensure that the `gganimate` and the `gifski` packages are installed also.

The verification package is described in some detail in its manual which is available from https://cran.r-project.org/web/packages/verification/verification.pdf.

Preliminaries

First, we empty the R environment of any existing variables and we load all required libraries.

```
rm(list=objects())
library(verification)
library(lubridate)
library(dplyr)
library(ggplot2)
library(reshape2)
options('max.print'=15)
```

Read data

We read the data from file. The various files contain readily paired forecast and observation data. Three files are read: deterministic forecasts ensemble forecasts and probabilistic forecasts, respectively. All files contain pairs of forecasts and their corresponding observations. The probabilistic forecast contains probabilities of the event that the future streamflow will exceed the value of 200 m33/s. Upon reading the file, the 'character' date/time columns are converted into a format that R interprets as a date/time.

```
pairs_det  <- read.csv('H-MS-SINT.det', as.is=T) %>% mutate(validtime=ymd_hms(validtime))
pairs_ens  <- read.csv('H-MS-SINT.ens', as.is=T) %>% mutate(validtime=ymd_hms(validtime))
pairs_prob <- read.csv('H-MS-SINT.prob',as.is=T) %>% mutate(validtime=ymd_hms(validtime))
head(pairs_det)
##             validtime leadtime     fcst    obs
## 1 2008-01-01 12:00:00        0 238.1200 238.12
## 2 2008-01-01 15:00:00        3 235.9806 236.11
## 3 2008-01-01 18:00:00        6 216.9600 244.53
##  [ reached 'max' / getOption("max.print") -- omitted 3 rows ]
head(pairs_ens)
##      validtime leadtime obs X01 X02 X03 X04 X05 X06 X07 X08 X09 X10 X11 X12 X13
##      X14 X15 X16
##  [ reached 'max' / getOption("max.print") -- omitted 6 rows ]
head(pairs_prob)
##             validtime leadtime fcst obs
## 1 2008-01-01 12:00:00        0 0.94   1
## 2 2008-01-01 15:00:00        3 0.94   1
## 3 2008-01-01 18:00:00        6 0.94   1
##  [ reached 'max' / getOption("max.print") -- omitted 3 rows ]
```

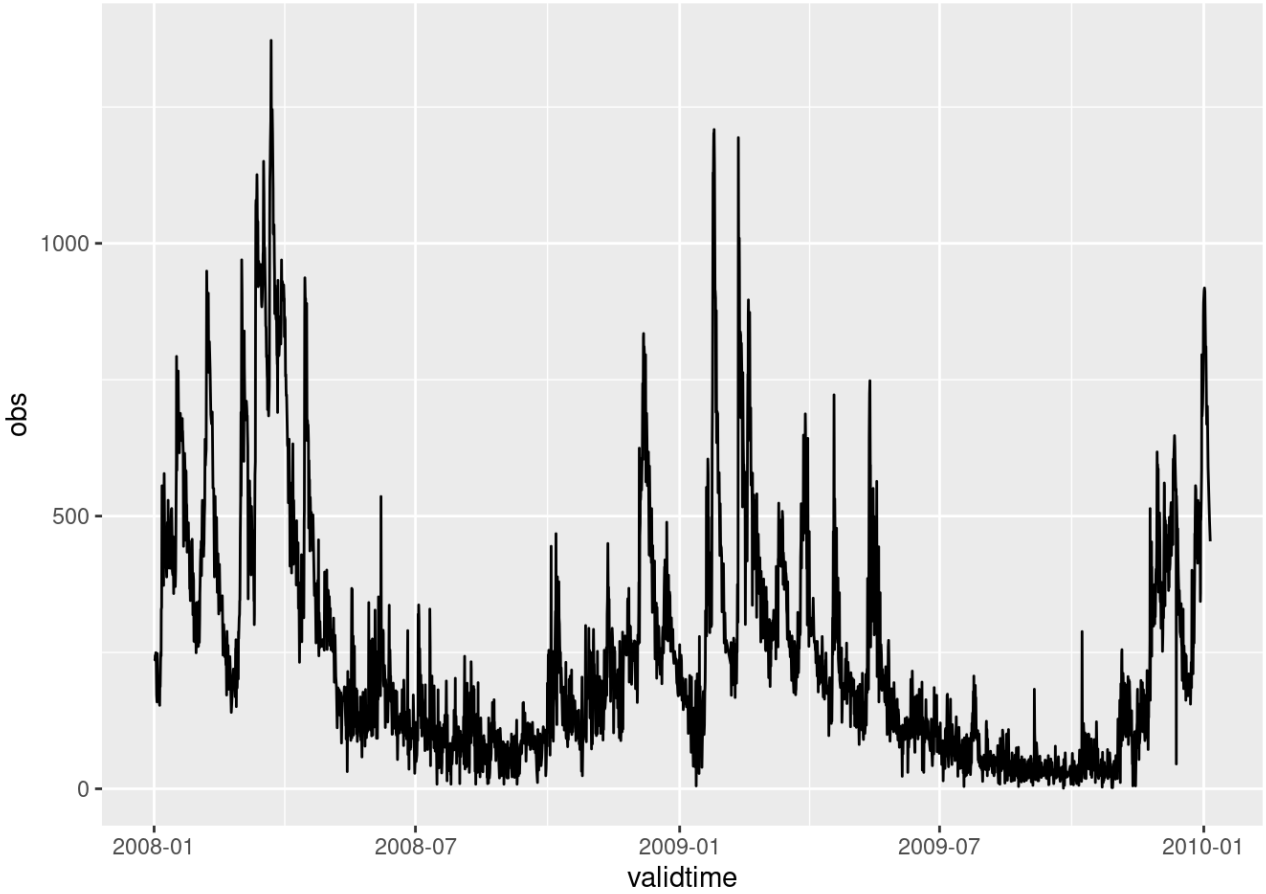
The R `verification` package does not directly use ensemble data. In the present example, it is imported for visualization only. Visualization using `ggplot` is much easier when the data is 'tidy' hence the dataframe is molten.

```
pairs_ens <- pairs_ens %>% melt(id.vars=c('validtime','leadtime','obs'),variable.name='me
mber',value.name='fcst')
head(pairs_ens)
##             validtime leadtime    obs member   fcst
## 1 2008-01-01 12:00:00        0 238.12    X01 238.12
## 2 2008-01-01 15:00:00        3 236.11    X01 235.98
## 3 2008-01-01 18:00:00        6 244.53    X01 216.95
##  [ reached 'max' / getOption("max.print") -- omitted 3 rows ]
```

Plot raw data

We plot observed values. For that, we create a new object `obs`:

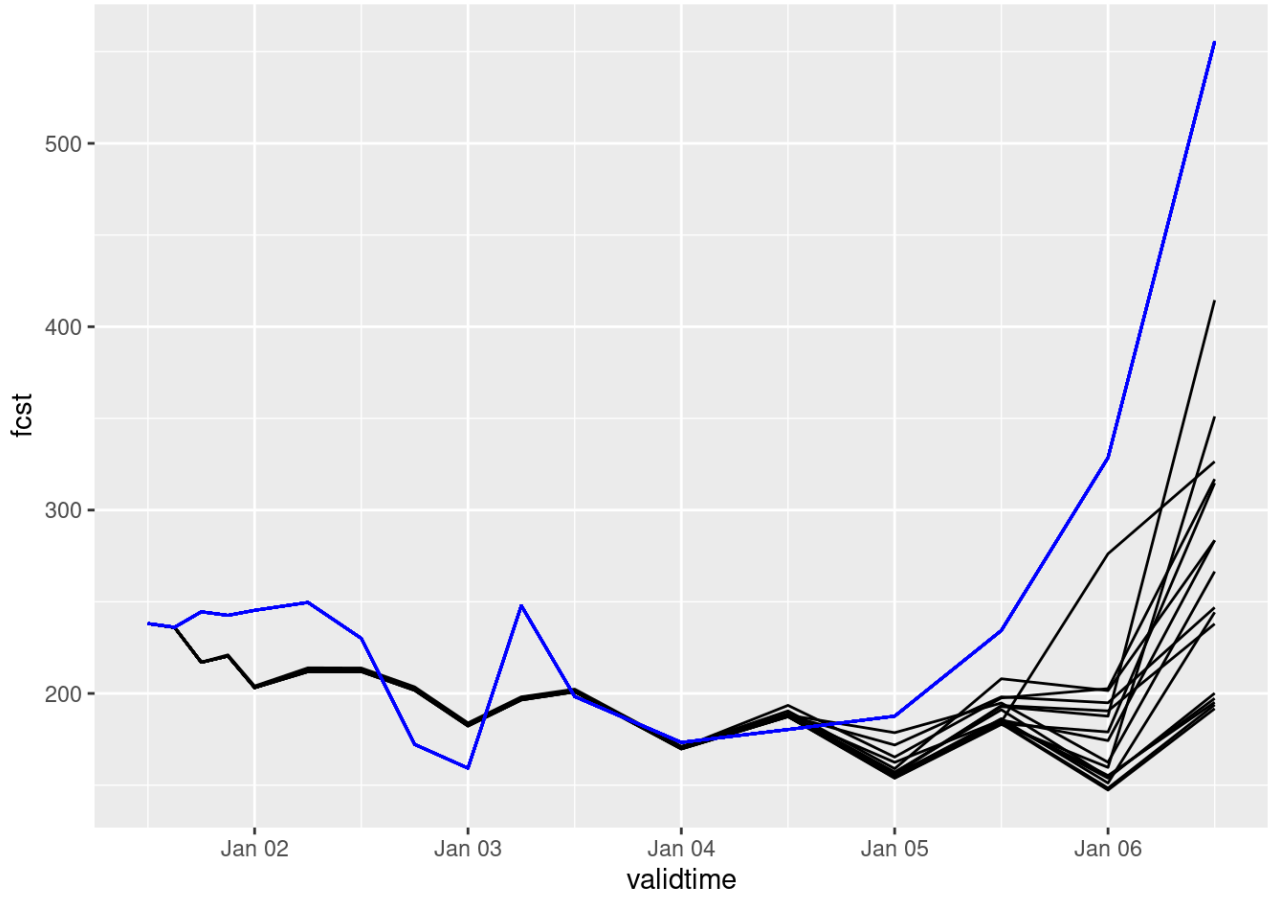
```
obs <- pairs_det %>% select(validtime, obs) %>% distinct()
ggplot(data=obs, aes(x=validtime,y=obs)) + geom_line() + scale_x_datetime()
```



Then we plot a hydrograph that combines forecast and observations. The data is filtered to show the January 1, 2008, 12Z forecast only.

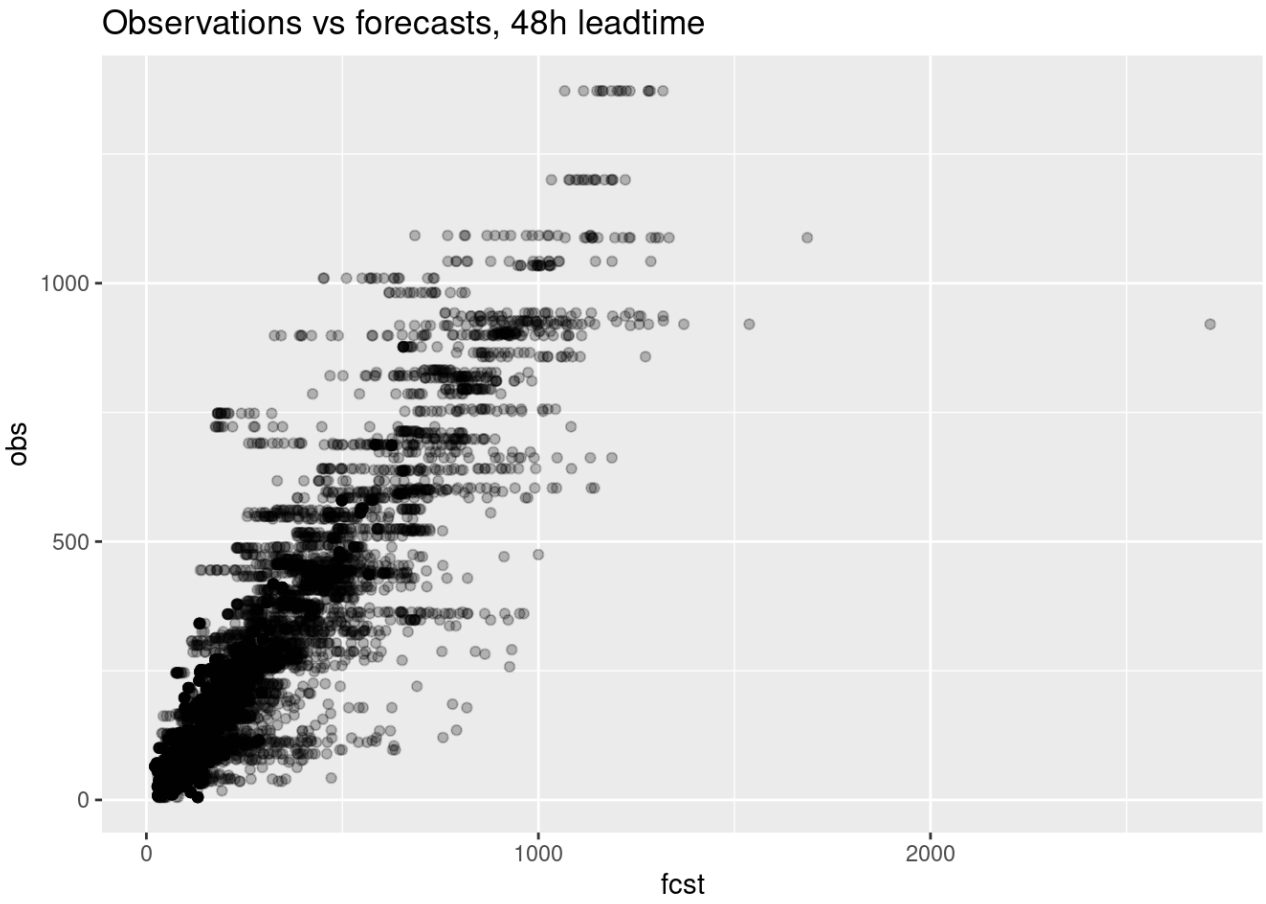
```
my_data <- pairs_ens %>% filter(validtime - hours(leadtime) == ymd_hm('200801011200'))


ggplot( data=my_data,aes(x=validtime, y=fcst, group=interaction(member))) +
  geom_line() + geom_line(aes(x=validtime, y=obs), col='blue')
```



Finally, we plot a scatter diagram.

```
ggplot(data=pairs_ens %>% filter(leadtime == 48), aes(x=fcst, y=obs)) +
  geom_point(alpha = 0.25) +
  labs(title = paste0('Observations vs forecasts, 48h leadtime')) +
  theme(legend.position = "none")
```

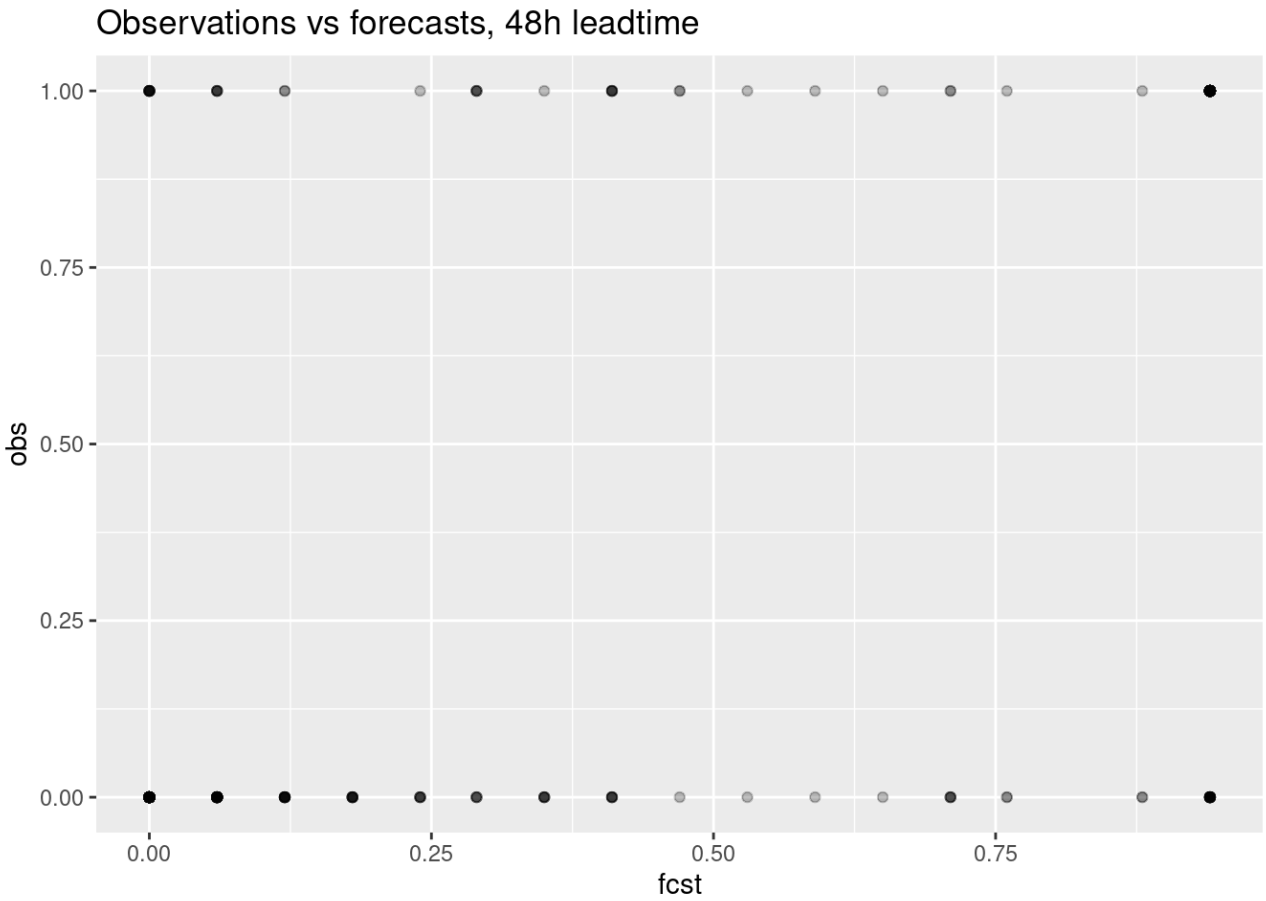


One could have opted to produce an animation instead.

```
# Not run
library('gganimate')
library('gifski')
ggplot(data=pairs_ens, aes(x=fcst, y=obs)) + geom_point(alpha = 0.25) +
  transition_states(leadtime, transition_length = 2, state_length = 1) + enter_fade() + e
xit_shrink() + ease_aes('sine-in-out') +
  labs(title = paste0('Observations vs forecasts, {closest_state}h leadtime')) + theme(le
gend.position = "none")
```

Below is a scatter plot of the probabilistic forecasts. Do you think this is a helpful plot?

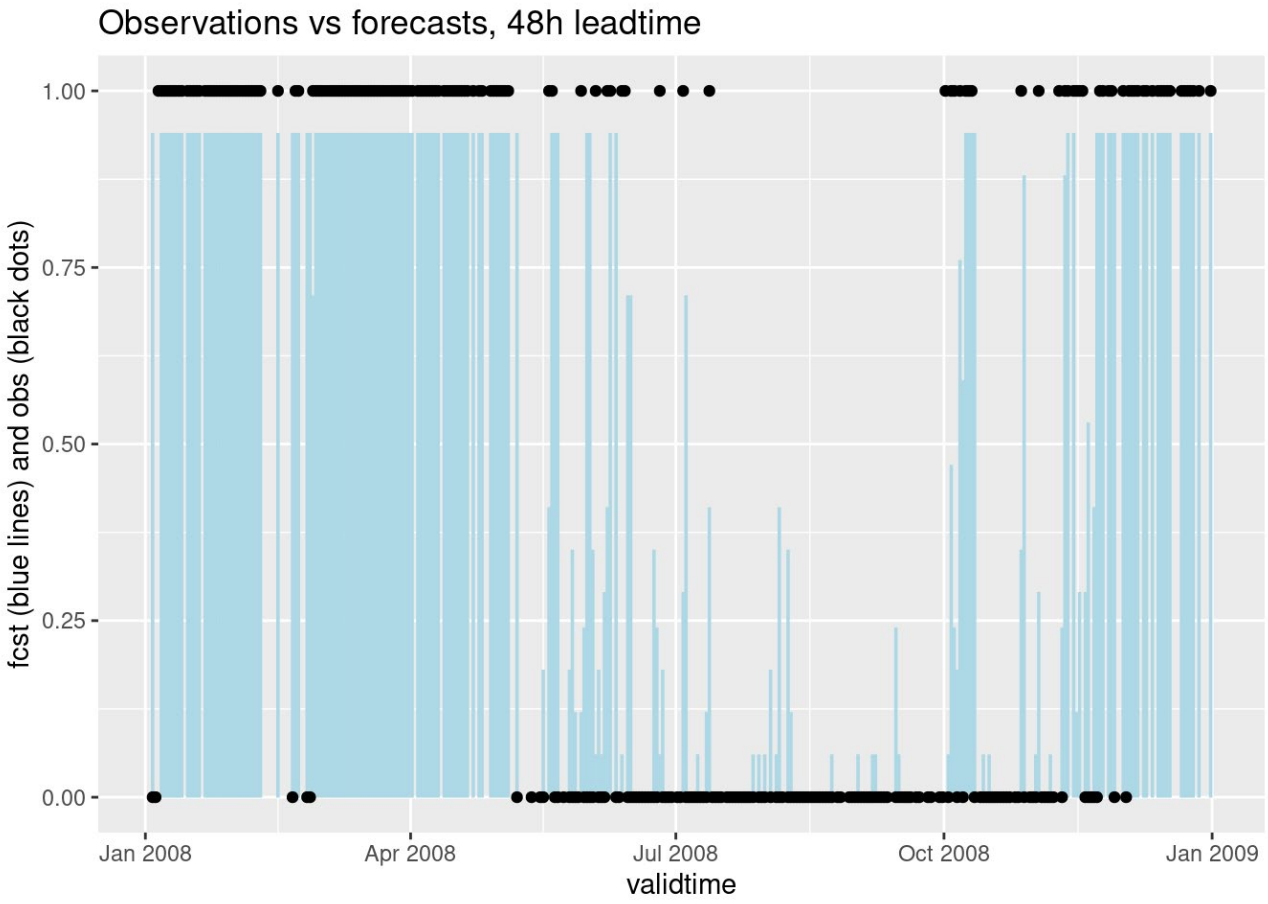
```
ggplot(data=pairs_prob %>% filter(leadtime == 48), aes(x=fcst, y=obs)) +
  geom_point(alpha = 0.25) +
  labs(title = paste0('Observations vs forecasts, 48h leadtime')) +
  theme(legend.position = "none")
```



We plot a timeseries of both the forecasts and the observations. This is a little more informative than above scatter plot.

```
my_data <- pairs_prob %>% filter(leadtime == 48, year(validtime)==2008)


ggplot(data=my_data, aes(x=validtime)) +
  geom_col(aes(y=fcst), colour = 'lightblue', fill = 'lightblue', size=0.15) +
  geom_point(aes(y=obs),  colour = 'black') +
  labs(title = paste0('Observations vs forecasts, 48h leadtime')) + #theme(legend.positio
n = "none") +
  ylab('fcst (blue lines) and obs (black dots)')
```



Verification of deterministic forecasts

The R **verification** package requires that a so-called *verification object* is created. This object contains some summary verification metrics and can be used to create various technical diagrams. We compute verification metrics for the 48-hour leadtime.

```
my_pairs <- pairs_det %>% filter(leadtime==48)
verify(obs=my_pairs$obs, pred=my_pairs$fcst, frcst.type='cont', obs.type='cont', show=F)
```

```
## $baseline.tf
## [1] FALSE
##
## $MAE
## [1] 55.00435
##
## $MSE
## [1] 6866.742
##
## $ME
## [1] 13.36633
##
## $MSE.baseline
## [1] 53745.48
```

```
##
## $MSE.pers
## [1] 11271.81
##
## $SS.baseline
## [1] 0.8722359
##
## $obs
##  [1] 198.31 180.33 234.33 555.59 373.76 406.16 421.17 443.34 441.78 429.20
## [11] 468.00 380.70 596.57 662.14 641.07
##  [ reached getOption("max.print") -- omitted 585 entries ]
##
## $pred
##  [1] 201.2594 183.2500 181.5188 395.0256 394.8388 489.0456 453.7744 471.8775
##  [9] 423.6919 592.8506 505.5431 380.2556 620.4394 879.1194 869.0400
##  [ reached getOption("max.print") -- omitted 585 entries ]
##
## $baseline
## [1] 253.1124
##
## attr(,"class")
## [1] "verify"    "cont.cont"
```

Verification of probabilistic forecasts

Then we do the same for the probabilistic forecasts:

```
my_pairs <- pairs_prob %>% filter(leadtime==48)

verify(obs=my_pairs$obs, pred=my_pairs$fcst, frcst.type='prob', obs.type='binary', show=F
)
```

```
## $baseline.tf
## [1] FALSE
##
## $bs
## [1] 0.0802
##
## $bs.baseline
## [1] 0.2473306
##
## $ss
## [1] 0.6757376
##
```

```
## $bs.reliability
## [1] 0.005193238
##
## $bs.resol
## [1] 0.1723238
##
## $bs.uncert
## [1] 0.2473306
##
## $y.i
##  [1] 0.05 0.15 0.25 0.35 0.45 0.55 0.65 0.75 0.85 0.95
##
## $obar.i
##  [1] 0.05415162 0.09523810 0.35714286 0.16666667 0.53846154 0.50000000
##  [7] 0.50000000 0.33333333 0.33333333 0.92430279
##
## $prob.y
##  [1] 0.461666667 0.035000000 0.023333333 0.010000000 0.021666667 0.006666667
##  [7] 0.003333333 0.015000000 0.005000000 0.418333333
##
## $obar
## [1] 0.4483333
##
## $thres
##  [1] 0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0
##
## $check
## [1] 0.0802
##
## $bins
## [1] TRUE
##
## $obs
##  [1] 0 0 1 1 1 1 1 1 1 1 1 1 1 1 1
##  [ reached getOption("max.print") -- omitted 585 entries ]
##
## $pred
##  [1] 0.94 0.00 0.00 0.94 0.94 0.94 0.94 0.94 0.94 0.94 0.94 0.94 0.94 0.94 0.94
##  [ reached getOption("max.print") -- omitted 585 entries ]
##
```

```
## attr(,"class")
## [1] "verify"    "prob.bin"
```

We create a new variable that contains the verification object; this allows for reusing the object for plotting various technical diagrams.

```
verification_object <- verify(obs=my_pairs$obs, pred=my_pairs$fcst, frcst.type='prob', obs.type='binary', show=F, thresholds = seq(-0.05,1.05,0.1))
```

Brier's probability score

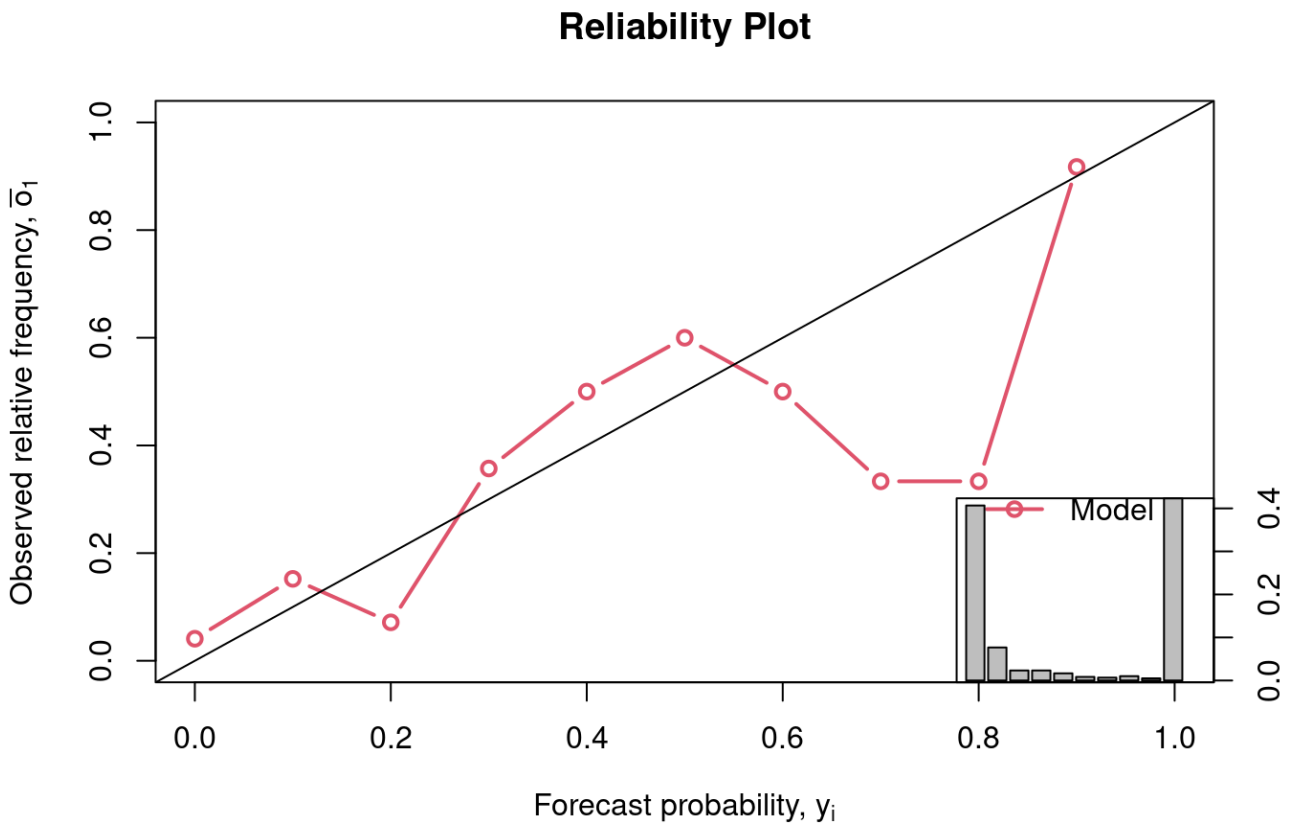
From the 'probabilistic' verification object, we can extract the value of various summary metrics including the Brier score and its decomposition:

```
verification_object$bs
## [1] 0.08028333
verification_object$bs.reliability
## [1] 0.004230695
verification_object$bs.resol
## [1] 0.1712779
verification_object$bs.uncert
## [1] 0.2473306
```

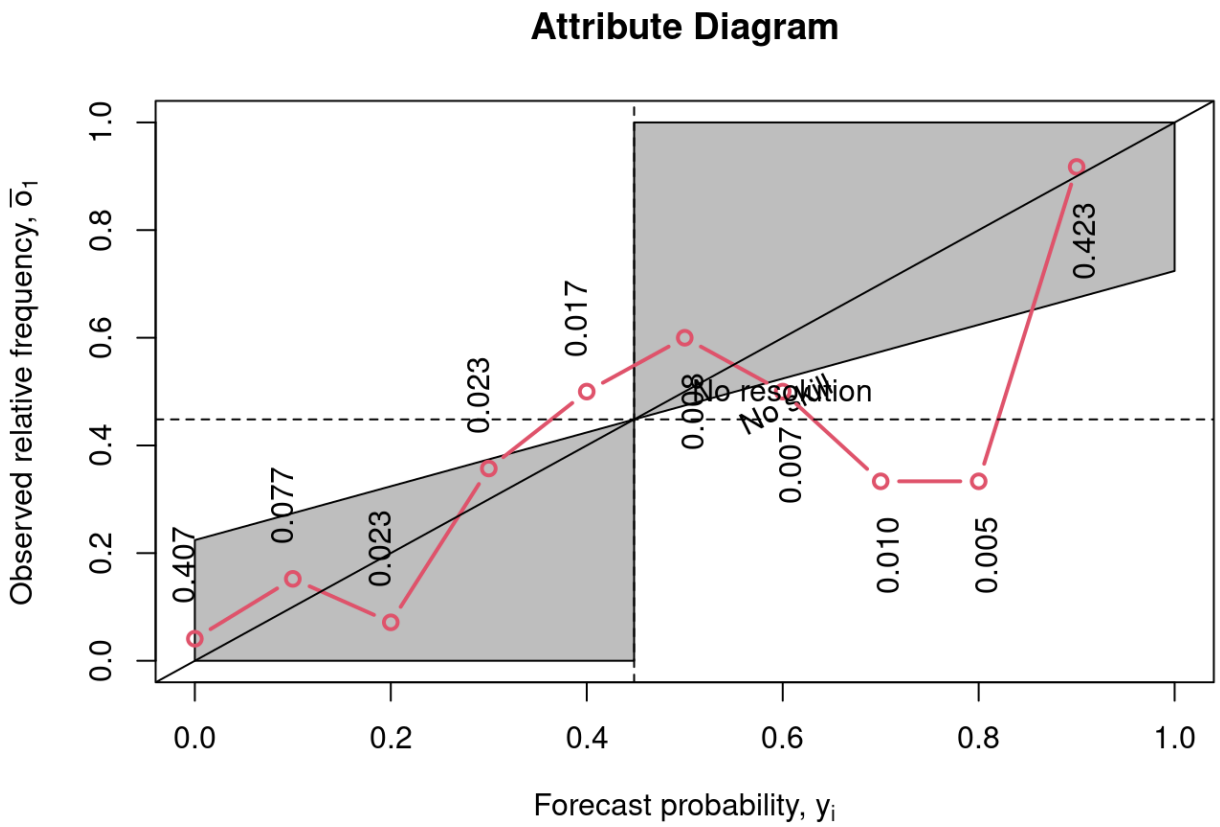
Technical diagrams

The 'probabilistic' verification object allows us to easily plot various technical diagrams such as the *reliability diagram*, the *attribute diagram* and the *relative operating characteristic*:

```
reliability.plot(verification_object)
```

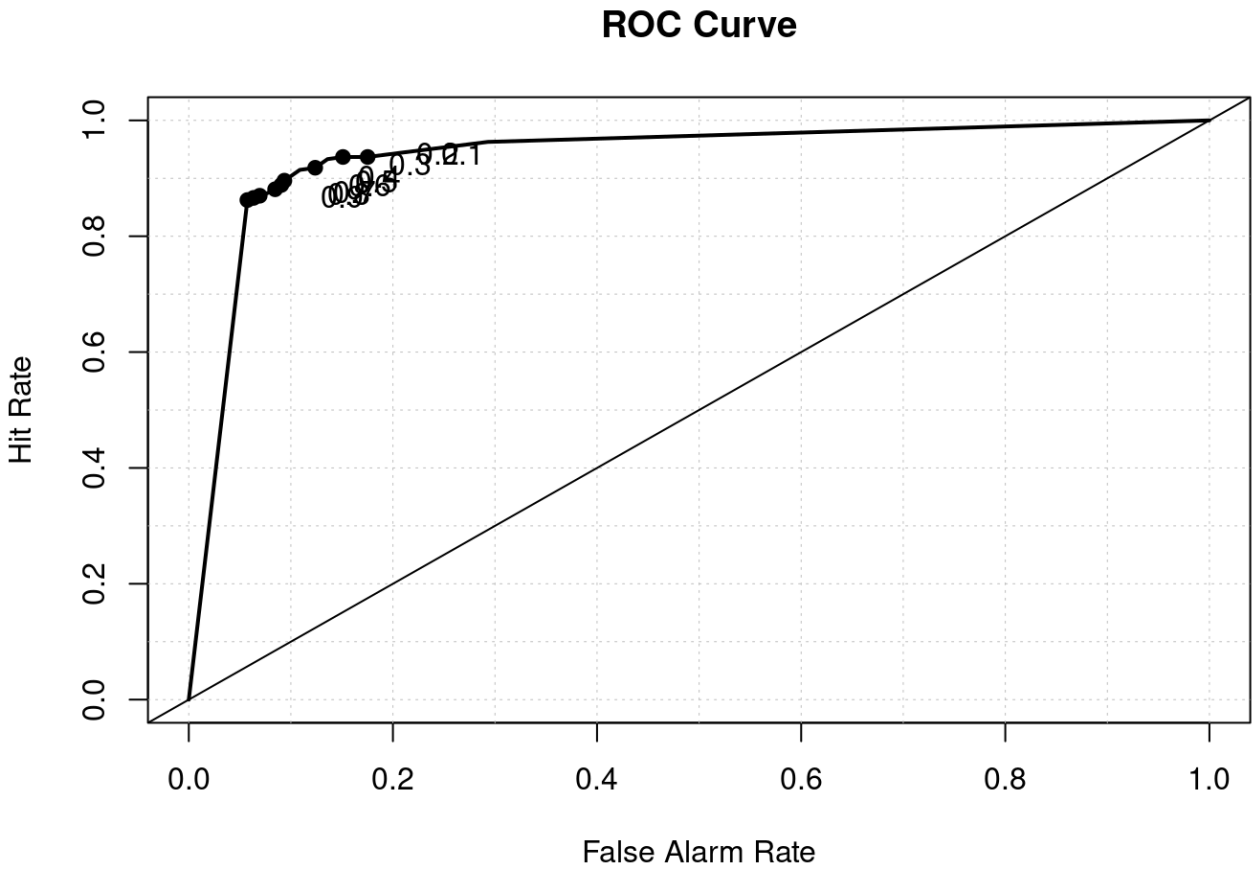



```
attribute(verification_object)
```




```
## NULL
```

```
roc.plot(verification_object)
```



A.8 Example 6 – Computational example: the verif package

The present example uses the verif forecast verification tool to assess the quality of a set of streamflow ensemble forecasts for a location on River Meuse. The tool can be downloaded from its portal at https://github.com/WFRT/verif. The portal includes installation instructions, documentations and examples. In addition, there is a wiki which describes the tool's features and constitutes a *de facto* manual: https://github.com/WFRT/verif/wiki.

The verif tool can be used as a Python module as well as a command line tool. In the present example, the verif package version 1.2.3 is used as a module within Python.

The tool includes many more options than those highlighted in the present document hence we encourage you to explore the tool's web portal and the wiki in more detail.

Preliminaries

First, the required libraries are loaded.

In [13]:

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import verif.data
import verif.input
import verif.interval
import verif.metric
```

Data

The verif tool allows for verification of deterministic and probabilistic forecasts. As the Meuse forecasts constitute ensemble forecasts, these have to be re-cast in either deterministic or probabilistic form. This was done prior the verification exercise. verif comes with a script that can do this; see https://github.com/WFRT/verif/wiki/Useful-scripts for details.

The wiki pages on the verif web portal outline how input data must be formatted: https://github.com/WFRT/verif/wiki/Arranging-my-own-data. For the present example, we have prepared two data files. One contains the deterministic forecasts and their verifying observations; the other the ensemble and the probabilistic forecasts and their verifying observations. The files contain data for a single location only, but note that the tool allows for including data for multiple locations.

The data files are structured as shown below. Note that for the purpose of visualization in the present document, the file contents are only read partially.

In [14]:
```
det_file = open('H-MS-SINT.det').read(500)
print(det_file)
```

```
# variable: Streamflow rate
# units: $m^3/s$
date hour leadtime obs fcst
20080101 12 0 238.12 238.12
20080101 12 3 236.11 235.980625
20080101 12 6 244.53 216.96
20080101 12 9 242.59 220.44875
20080101 12 12 245.33 203.230625
20080101 12 18 249.64 212.48
20080101 12 24 230.03 212.38375
20080101 12 30 172.21 202.2125
20080101 12 36 159.24 182.690625
20080101 12 42 247.9 196.921875
20080101 12 48 198.31 201.259375
20080101 12 60 173.3 170.150625
20080101 12 72 180.33 188.4
20080101 12 84 187.61 159
```

In [15]:
```
ens_file = open('H-MS-SINT.ens').read(1000)
print(ens_file)
```

```
# variable: Streamflow rate
# units: $m^3/s$
date hour leadtime obs 01 02 03 04 05 06 07 08 09 10 11 12 13 14 15 16 p100 p200 p400
p800 q0 q0.1 q0.2 q0.3 q0.4 q0.5 q0.6 q0.7 q0.8 q0.9 q1
20080101 12 0 238.12 238.12 238.12 238.12 238.12 238.12 238.12 238.12 238.12 238.12 238.1
2 238.12 238.12 238.12 238.12 238.12 238.12 0.0588235294117647 0.0588235294117647 1 1 238
.12 238.12 238.12 238.12 238.12 238.12 238.12 238.12 238.12 238.12 238.12
20080101 15 3 236.11 235.98 235.98 235.98 235.98 235.98 235.98 235.98 235.98 235.98 2
35.99 235.98 235.98 235.98 235.98 235.98 235.98 0.0588235294117647 0.0588235294117647 1 1
235.98 235.98 235.98 235.98 235.98 235.98 235.98 235.98 235.98 235.98 235.99
20080101 18 6 244.53 216.95 216.95 216.95 216.95 216.95 216.95 217.08 216.95 216.95 2
16.96 216.95 216.95 216.95 216.95 216.97 216.95 0.0588235294117647 0.0588235294117647 1 1
216.95 216.95 216.95 216.95 216.95 216.95 216.95 216.95 216.95 216.965 217.08
20080101 21 9 242.59 220.4 220.4 220.4 220.4 220.4 220.4
```

The contents of the H-MS-SINT.det and H-MS-SINT.ens files are used to create verification objects of the verif.data.Data type. As for some purposes, we will only want to plot and/or compute data for a specific lead time, a second verification object (_48) is created.

In [16]:
```
pairs_det = verif.input.get_input("H-MS-SINT.det")
data_det  = verif.data.Data([pairs_det])
data_det_48h = verif.data.Data([pairs_det], leadtimes=[48])
data_det_20080101 = verif.data.Data([pairs_det],dates=[20080101])
```

```
print(data_det)
```

```
<verif.data.Data object at 0x000002B2386A1C10>
```

In [17]:

```
pairs_ens = verif.input.get_input("H-MS-SINT.ens")
data_ens  = verif.data.Data([pairs_ens])
data_ens_48h = verif.data.Data([pairs_ens], leadtimes=[48])
print(data_ens)
```

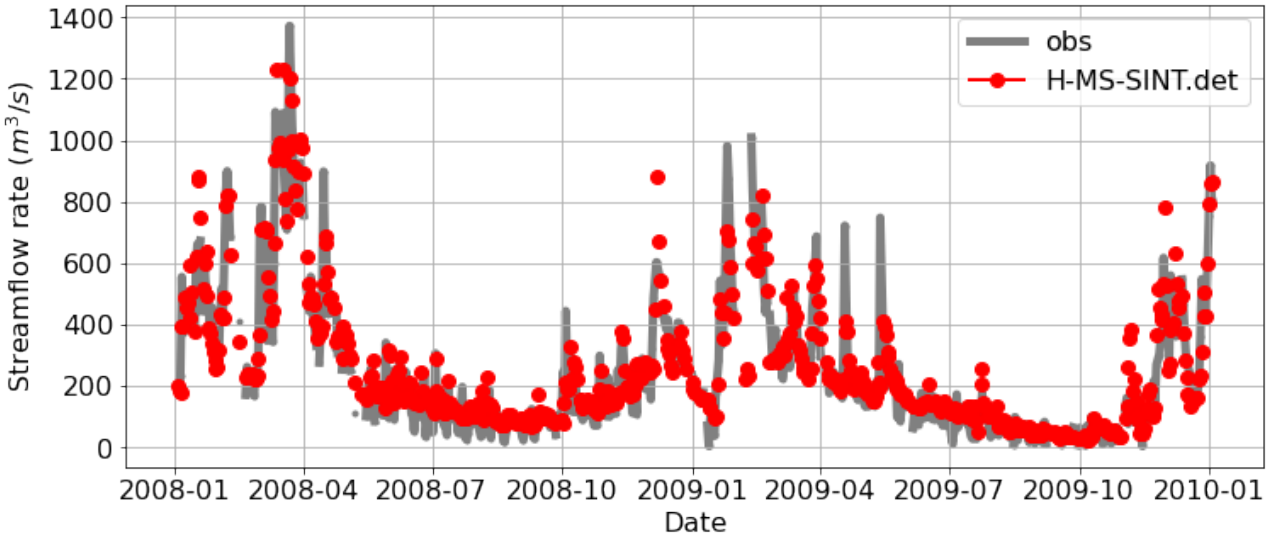
```
<verif.data.Data object at 0x000002B2370F8400>
```

Exploration of raw data

From the verification objects, the raw data can be explored. verif allows for plotting long-term timeseries, a single forecast hydrograph and the verifying observations, and scatter plots. Note that this cannot be done for ensemble forecasts.

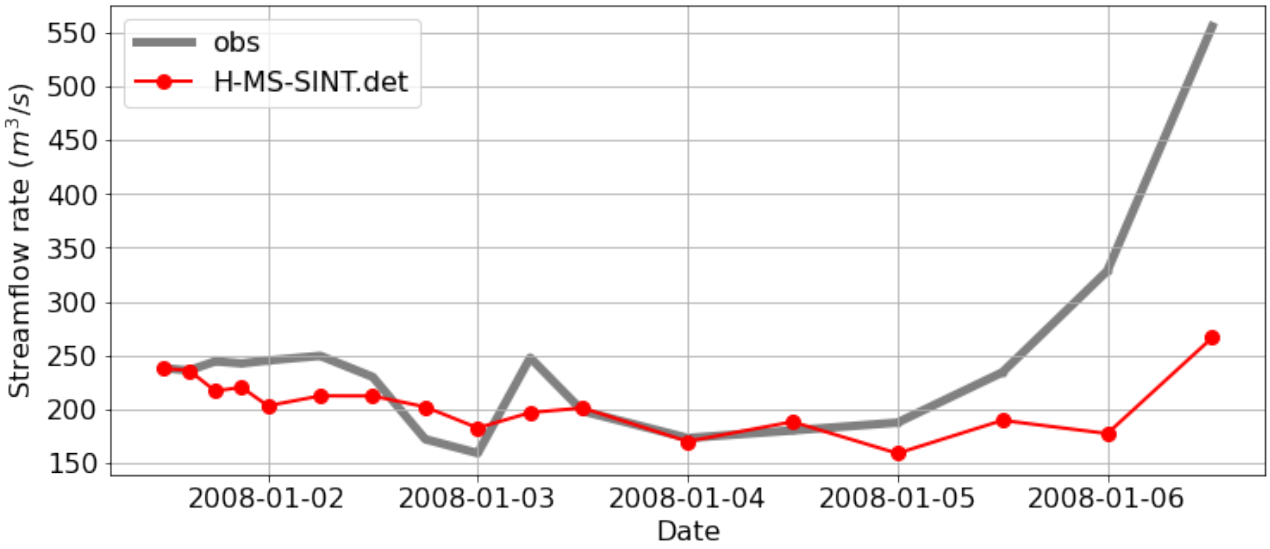
In [18]:

```
output = verif.output.TimeSeries()
output.figsize = [12,5]
output.plot(data_det_48h)
```



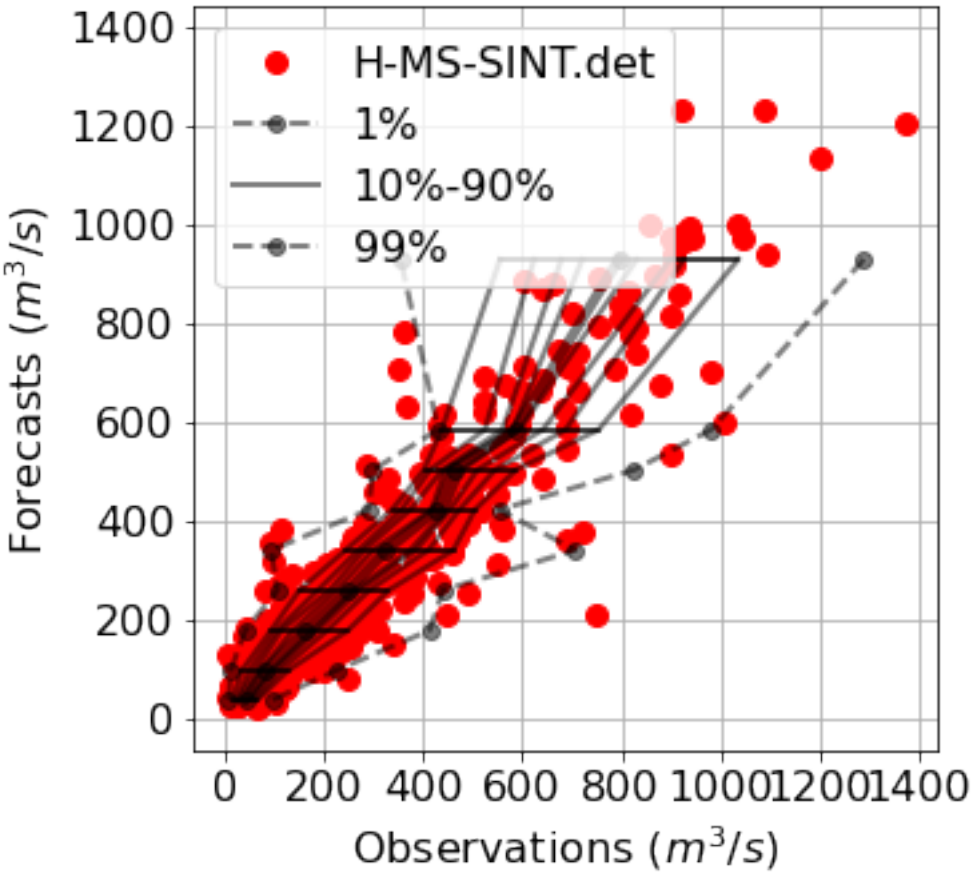
In [19]:

```
output = verif.output.TimeSeries()
output.figsize = [12,5]
output.plot(data_det_20080101)
#output.csv(data_det) #for generating text output
```



In [20]:

```
output = verif.output.Scatter()
output.plot(data_det_48h)
```



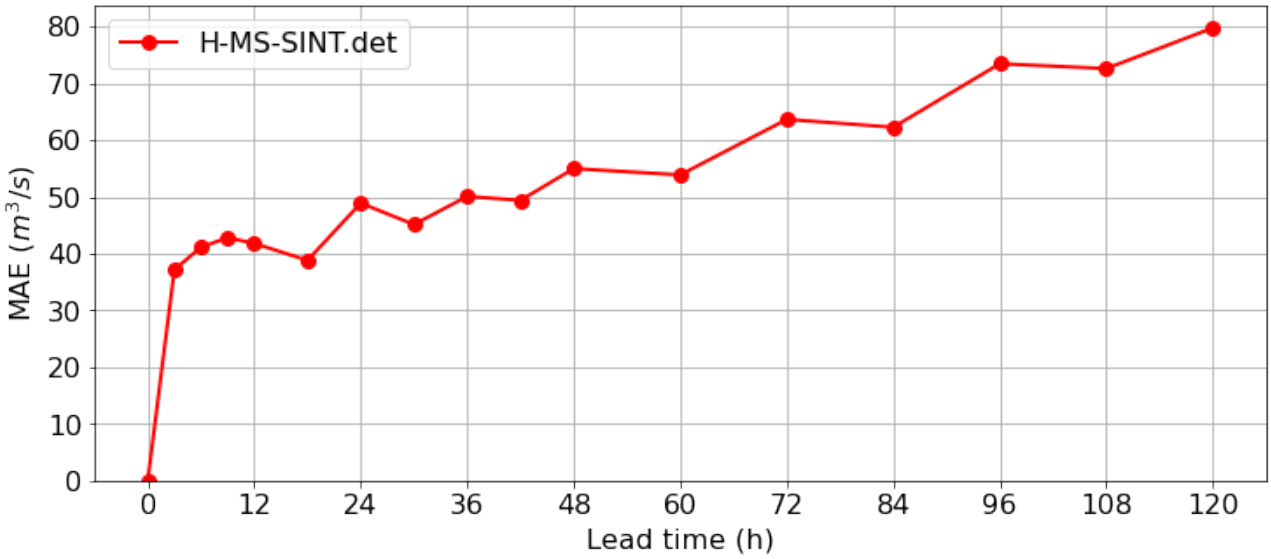
Verification of deterministic forecasts

verif allows for summarizing the quality of forecasts using a large number of summary metrics. These include bias, mean bias, mean absolute error and root mean squared error. A full list is shown at https://github.com/WFRT/verif/wiki/List-of-metrics. Here, the computation of mean absolute error (mae), the root mean squared error (rmse) and bias is demonstrated.

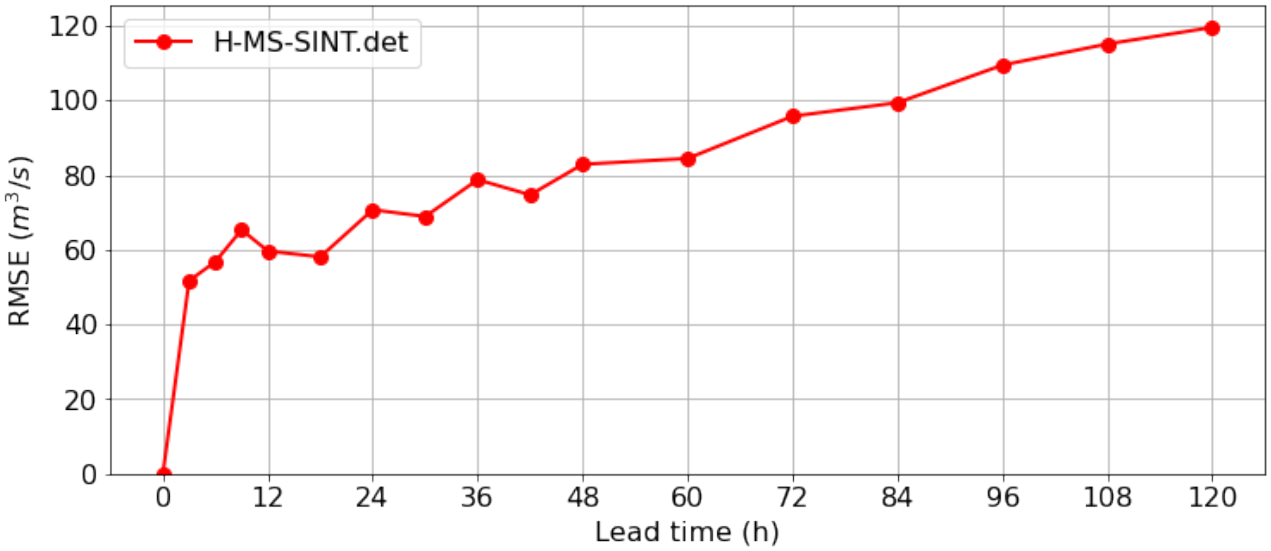
In [21]:

```
metrics = [verif.metric.Mae(), verif.metric.Rmse(), verif.metric.Bias()]
```

```
for metric in metrics:
    output = verif.output.Standard(metric)
    output.figsize = [12,5]
    output.plot(data_det)
    output.csv(data_det)
```

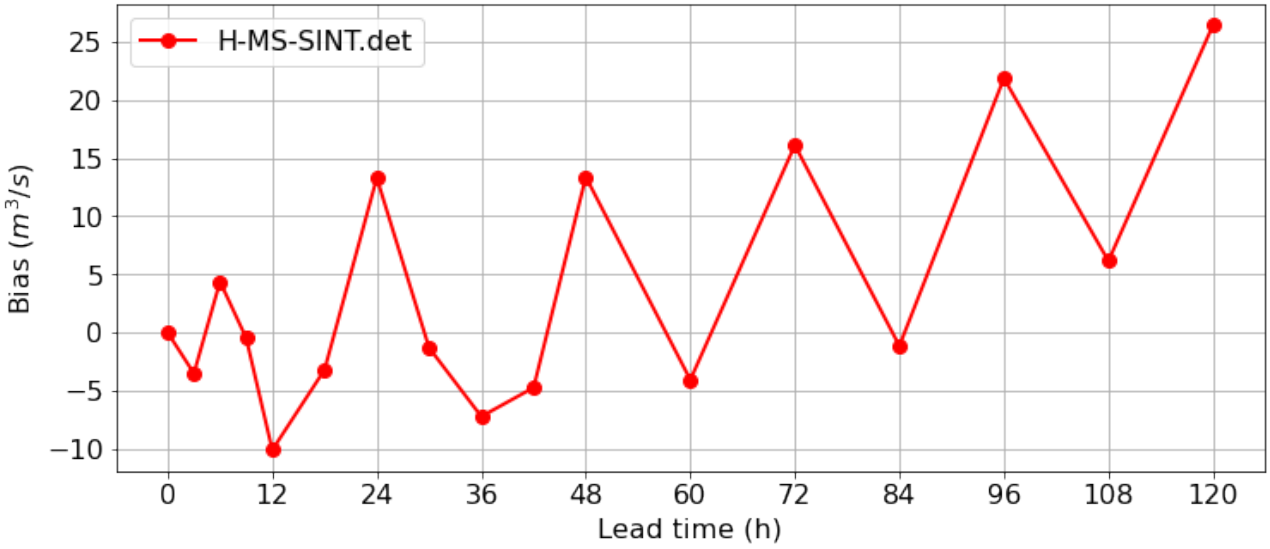



```
Leadtime,H-MS-SINT.det
0.0,0
3.0,37.1822
6.0,41.1568
9.0,42.831
12.0,41.8123
18.0,38.8361
24.0,48.9172
30.0,45.1477
36.0,50.1115
42.0,49.4071
48.0,55.0043
60.0,53.8959
72.0,63.6788
84.0,62.2885
96.0,73.4854
108.0,72.6281
120.0,79.7995
```




```
Leadtime,H-MS-SINT.det
0.0,0
3.0,51.6739
6.0,56.8027
```

```
9.0,65.3277
12.0,59.6484
18.0,58.0823
24.0,70.6843
30.0,68.8936
36.0,78.7351
42.0,74.677
48.0,82.8658
60.0,84.4118
72.0,95.7289
84.0,99.3433
96.0,109.408
108.0,115.059
120.0,119.503
```




```
Leadtime,H-MS-SINT.det
0.0,0
3.0,-3.49366
6.0,4.38072
9.0,-0.452529
12.0,-10.0647
18.0,-3.1836
24.0,13.2834
30.0,-1.29227
36.0,-7.19521
42.0,-4.71902
48.0,13.3663
60.0,-4.0438
72.0,16.0964
84.0,-1.13481
96.0,21.8324
108.0,6.15552
120.0,26.3862
```

Verification of probabilistic forecasts

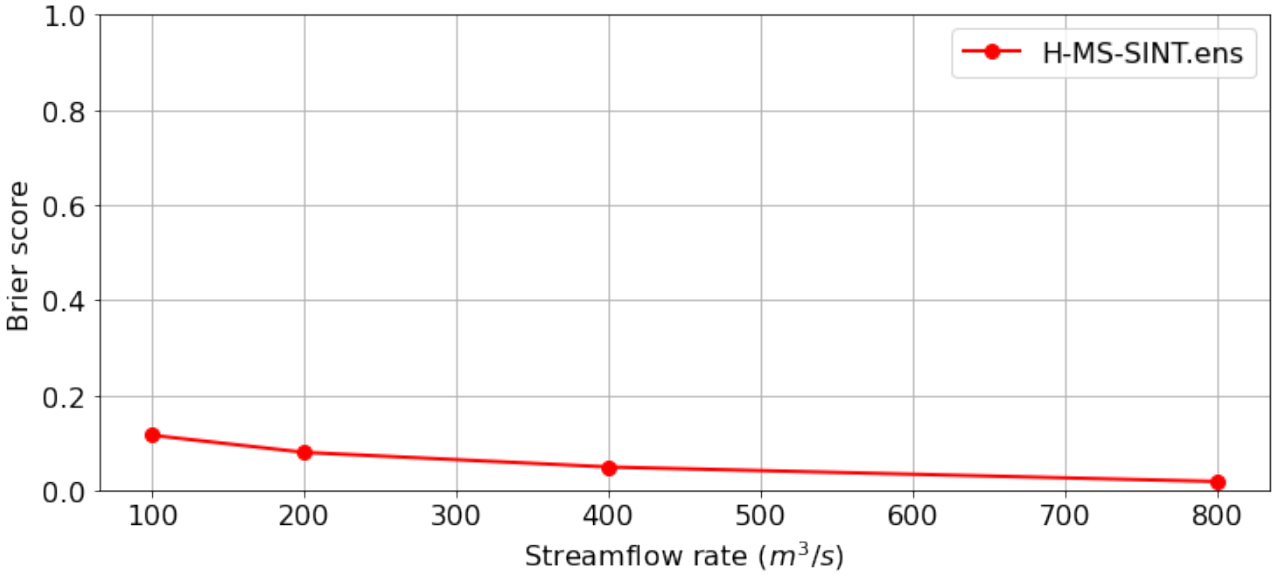
Various metrics that measure the quality of probabilistic forecasts can be computed. A full list is shown below the 'probabilistic' header at https://github.com/WFRT/verif/wiki/List-of-metrics. In the present document, Brier's probability score as well as various decompositions are computed.

Brier's probability score

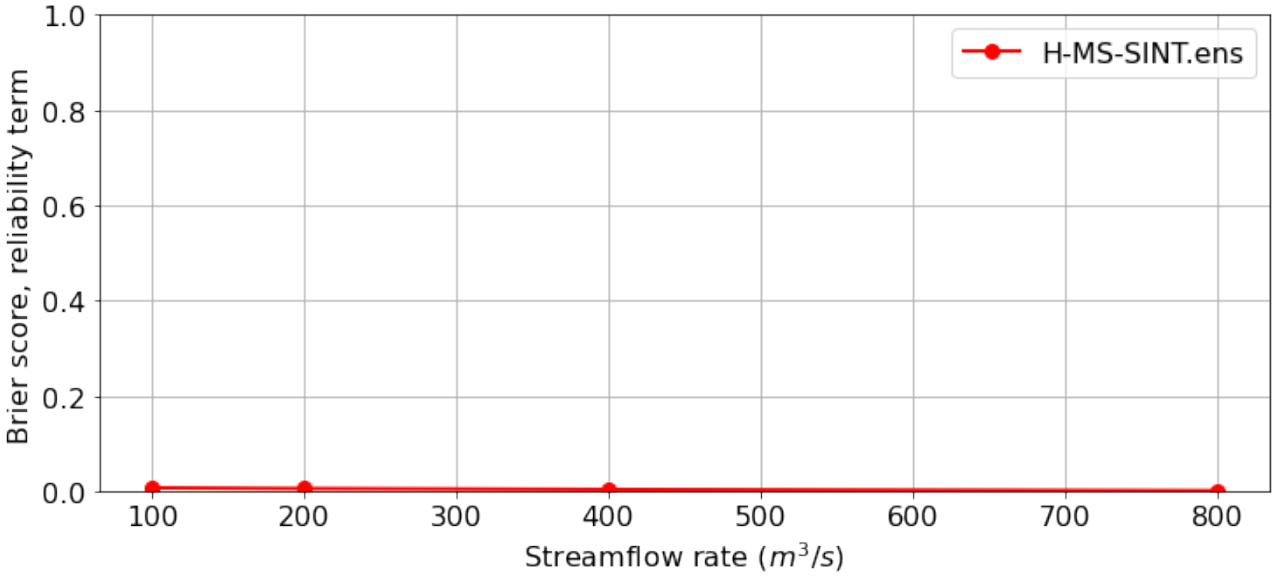
In [22]:

```
    metrics = [verif.metric.Bs(), verif.metric.BsRel(),verif.metric.BsRes(),
verif.metric.BsUnc()]
    for metric in metrics:
        output = verif.output.Standard(metric)
```

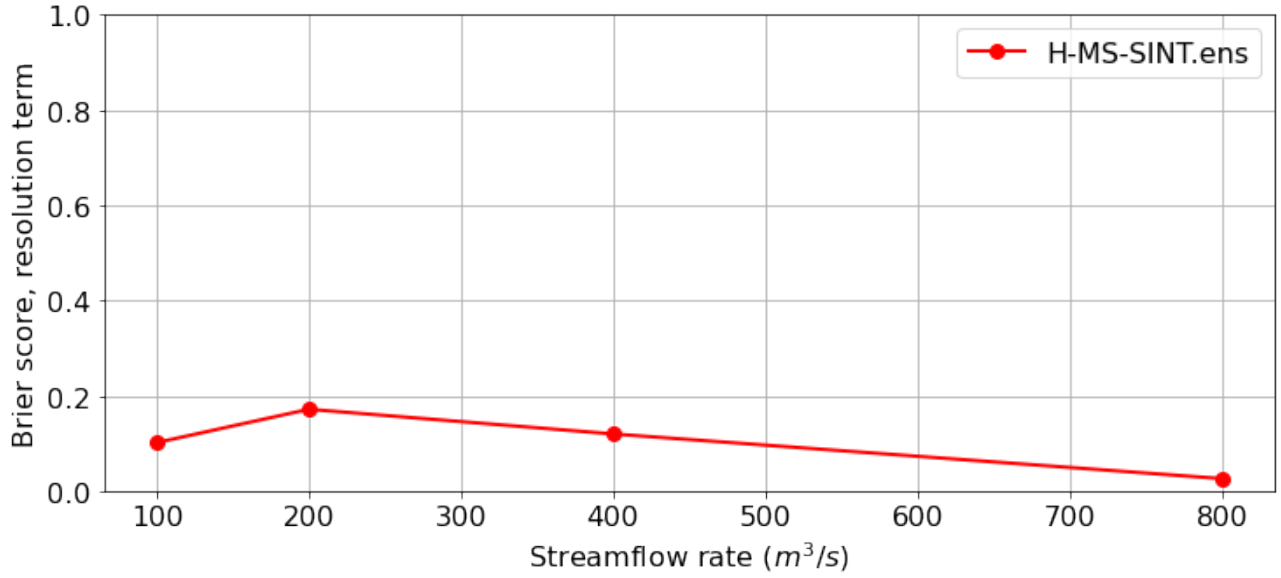
```
        output.thresholds = [100,200,400,800] #Note that these were the thresholds that
were included in the input .ens file
        output.figsize = [12,5]
        output.plot(data_ens_48h)
        output.csv(data_ens_48h)
```



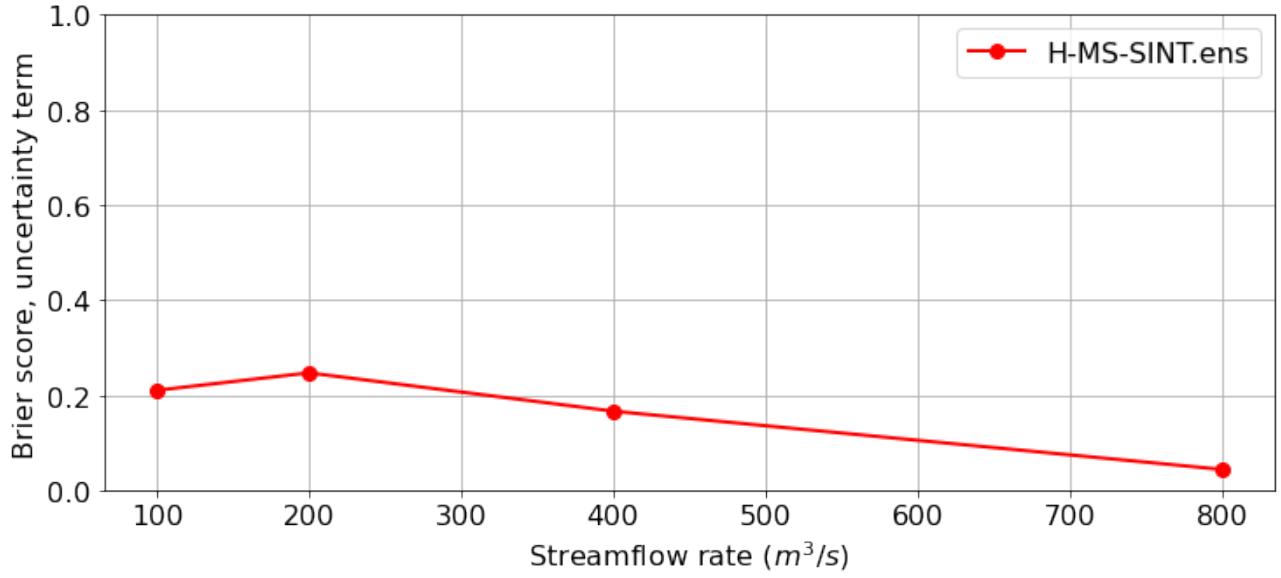

```
Threshold,H-MS-SINT.ens
100,0.116355
200,0.0804037
400,0.0496597
800,0.019158
```




```
Threshold,H-MS-SINT.ens
100,0.00777652
200,0.00617704
400,0.00414394
800,0.00170024
```




```
Threshold,H-MS-SINT.ens
100,0.102381
200,0.172324
400,0.120565
800,0.0268846
```



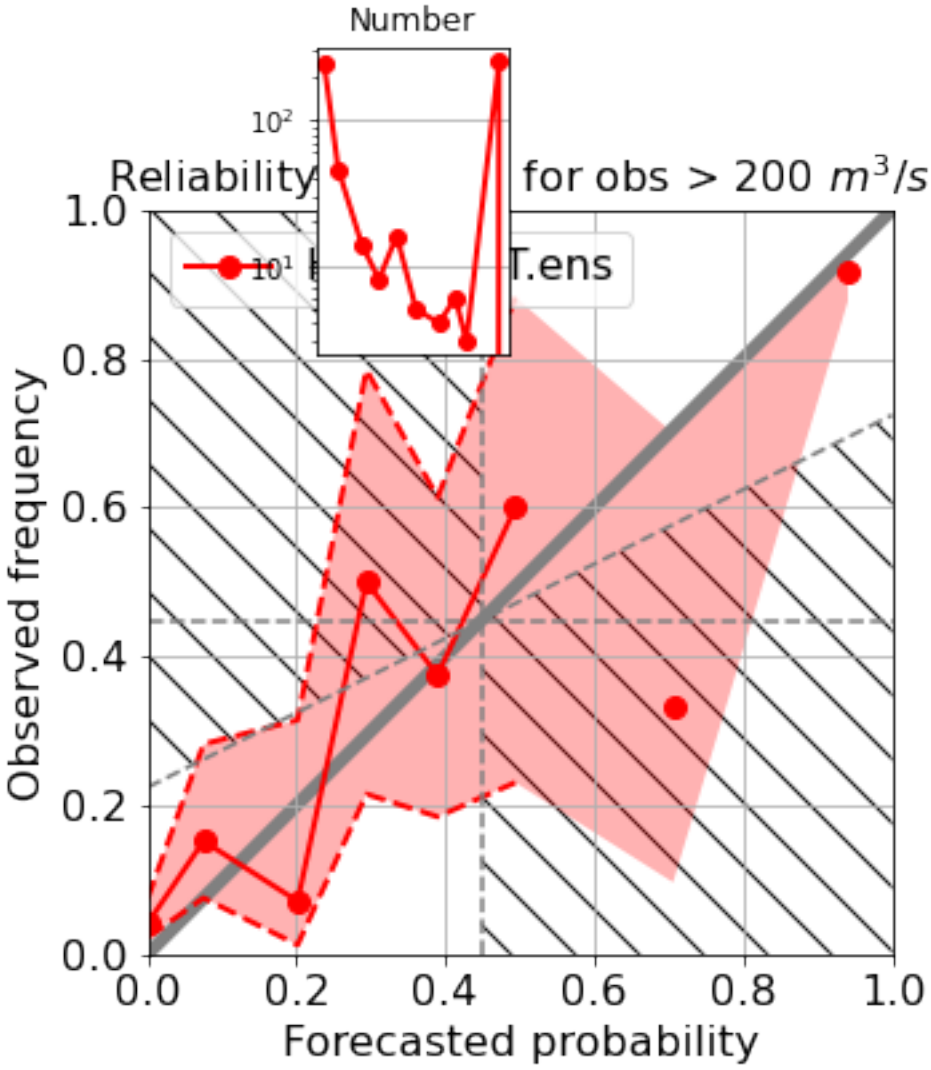

```
Threshold,H-MS-SINT.ens
100,0.210664
200,0.247331
400,0.166864
800,0.0444889
```

Technical diagrams

verif allows for creating various technical diagrams including the reliability diagram and the relative operating characteristic.

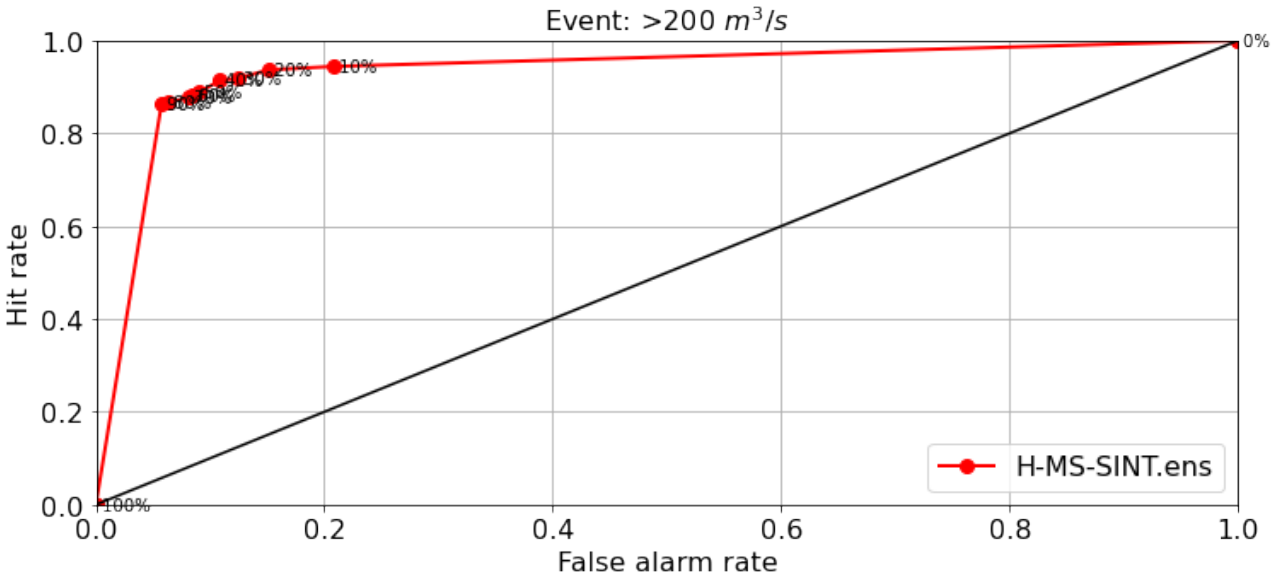
In [23]

```
output = verif.output.Reliability()
output.thresholds = [200]
output.plot(data_ens_48h)
```



In [24]:

```
output = verif.output.Roc()
output.thresholds = [200]
output.figsize = [12,5]
output.plot(data_ens_48h)
```



A.9 Example 7 – Compare accuracy, reliability and sharpness of two ensemble streamflow forecasting systems in a stream with zero flow

Setup

We first must load data, specify the forecast evaluation period and select the streamflow gauge to verify.

Add paths to data and code

First, we need to point to the data and functions needed to run this code.

```
sPath = matlab.desktop.editor.getActiveFilename;
indx = strfind(sPath,filesep);
addpath(genpath(sPath(1:indx(end-1)-1)));
```

Load forecasts and data

Set verification period

Because of the large volumes of data involved, for this example we verify forecasts issued for a single month (a total of 56 forecasts). The month includes zero and non-zero observations.

In general, it is desirable to assess as many retrospective forecasts as possible to ensure verification statistics are robust, so long as each forecast–observation pair is independent of other pairs. In this case we are calculating verification metrics for each lead time, and forecasts are issued every 12 hours. This means forecast–observations pairs are separated by at least 12 hours. It is likely that streamflow data 12 hours apart are correlated to some extent, meaning that some of the points will not be strictly independent. This mainly affects the statistical significance tests we perform for reliability. For simplicity, we ignore the assumption of independence throughout this example, and describe how this issue can be dealt with for significance tests of reliability when we describe those methods. Note also that as we are testing only a small subset of the forecasts Bennett et al. (2021) used, the results in this example differ somewhat from those presented in that study.

```
% The start and end dates of the verification period (inclusive)
stDate = datetime(2010,2,1);
enDate = datetime(2010,2,28,23,0,0);
```

Set gauge to be verified

Forecasts are generated at multiple sites, but we focus on only one site to keep this example simple.

```
stationId = 4;
```

Load data and forecasts from netCDF files

We store forecasts and data in netCDF files of our own specification. This specification allows us to store forecasts issued at multiple times and multiple locations in a single file. Forecasts are generated twice daily at the hourly timestep, and feature a large number of ensemble members (n = 1 000). The specification is documented at https://github.com/csiro-hydroinformatics/efts/blob/master/docs/netcdf_for_water_forecasting.md.

```
% Load observations
obsNc = "Brisbane_flow_utc_station.nc";
obsQ = readFctNc(obsNc,'q_obs');
```

```
% Load forecasts from method 1: the 'old' ERRIS method
oldFctNc = "Brisbane_GR4MUSKERRIS_7DAY_fct_Obsfrcng_xv2010_rstrcOn.nc";
oldFctQ = readFctNc(oldFctNc,'q_sim','start',stDate,'end',enDate);

% Load forecasts from method 2: the 'new' ERRIS method
newFctNc = "Brisbane_GR4MUSKMAERRIS_7DAY_fct_Obsfrcng_xv2010_s2W240genR1.nc";
newFctQ = readFctNc(newFctNc,'q_sim','start',stDate,'end',enDate);

% Get units of lead time (for use in generating climatology later)
leadTimeUnits = ncreadatt(newFctNc,'lead_time','units');
strInd = strfind(leadTimeUnits,' since');
leadTimeUnits = leadTimeUnits(1:strInd(1)-1);
```

Order observations for verification

This step orders observations to simplify the calculation of performance scores conditioned on lead time.

```
% Allocate memory
verObs = zeros(size(newFctQ.data,[1 4]))*nan;

% Order observations by forecast dates
for t = 1:size(newFctQ.timeUTC)
   [~,tIndx] = ismember(newFctQ.leadTimeUTC(:,t),obsQ.timeUTC);
   sIndx = obsQ.stationId==stationId;
   verObs(:,t) = obsQ.data(sIndx,tIndx)';
end
```

Develop a cross-validated climatology forecast

We develop a climatology by sampling sequences from historical observations at ordinal dates close to the forecast dates. In this case we have enough observations to generate a large ensemble. It is possible to verify against a smaller ensemble, if need be, but care needs to be taken when comparing forecasts with different ensemble sizes.

We now sample observed streamflows from similar ordinal dates but from years other than that for which the forecast is generated. This can take around 30 seconds for 1 month of forecasts.

```
% Allocate memory
climFct = zeros(size(newFctQ.data,[1 3 4]))*nan;

% Build climatology from observations
nanIndx = ~isnan(obsQ.data(obsQ.stationId==stationId,:));
tSt = find(nanIndx' & month(obsQ.timeUTC)==1 & day(obsQ.timeUTC)==1 ...
   & hour(obsQ.timeUTC)==0,1,"first");
tEnd = find(nanIndx' & month(obsQ.timeUTC)==12 & day(obsQ.timeUTC)==31 ...
   & hour(obsQ.timeUTC)==23,1,"last");
obsYrs = unique(year(obsQ.timeUTC(tSt:tEnd)));
fctYrs = unique(year(newFctQ.timeUTC));
climYrs = obsYrs(~ismember(obsYrs,fctYrs));
ensSize = length(newFctQ.realization);
```

```
noSampleDates = ceil(ensSize/length(climYrs));
noLeadTimes = size(newFctQ.leadTimeUTC,1);
sIndx = obsQ.stationId == stationId;

w = waitbar(0,'Sampling obs...','Name','Generate climatology forecast');
for f = 1:length(newFctQ.timeUTC)
   % Set up ensemble member counter
   waitbar(f/length(newFctQ.timeUTC));
   ensNo = 1;
   continueSampling = true;
   % Fill each ensemble member individually
   while ensNo<=ensSize
      % Loop through years from which climatology is to be built
      for y = 1:length(climYrs)
         yr = climYrs(y);
         if yr == climYrs(end) && ensNo<(ensSize-noSampleDates)
            % This is needed in the case of ensemble members knocked
            % out for null values
            noSampleDates = ensSize-ensNo+1;
         end
         stSampDate = max(datetime(yr,month(newFctQ.leadTimeUTC(1,f)),...
            day(newFctQ.leadTimeUTC(1,f)),hour(newFctQ.leadTimeUTC(1,f)),...
            minute(newFctQ.leadTimeUTC(1,f)),second(newFctQ.leadTimeUTC(1,f)))...
            -days(floor(noSampleDates/2)), obsQ.timeUTC(1));
         if yr == climYrs(end)
            switch leadTimeUnits
               case 'hours'
                  sampDates = min(stSampDate,obsQ.timeUTC(end) ...
                     - days(noSampleDates + ceil(days(hours(noLeadTimes)))));
               case 'days'
                  sampDates = min(stSampDate,obsQ.timeUTC(end) ...
                     - days(noSampleDates + noLeadTimes));
               case 'months'
                  sampDates = min(stSampDate,obsQ.timeUTC(end) ...
                     - (days(noSampleDates) + calMonths(noLeadTimes)));
            end
         end
         for t = 1:noSampleDates
            switch leadTimeUnits
               case 'hours'
                  sampDates = stSampDate+days(t-1):hours(1): ...
                     stSampDate+days(t-1)+hours(noLeadTimes-1);
               case 'days'
                  sampDates = stSampDate+days(t-1):days(1): ...
                     stSampDate+days(t-1)+days(noLeadTimes-1);
               case 'months'
                  sampDates = stSampDate+days(t-1):calMonths(1): ...
                     stSampDate+days(t-1)+calmonths(noLeadTimes-1);
            end
```

```
            [~,tIndx] = ismember(sampDates,obsQ.timeUTC);
            ensDat = obsQ.data(sIndx,tIndx);

            if any(isnan(ensDat))
               % We do not want climatology to have null values
               continue
            end

            climFct(:,ensNo,f) = ensDat;
            ensNo = ensNo+1;
            if ensNo>ensSize
               continueSampling = false;
               break
            end
            if climYrs(y) == climYrs(end) && t == noSampleDates
               warning('some ensemble members have not been filled')
               continueSampling = false;
            end

         end
         if ensNo>ensSize
            break
         end
      end
      if ~continueSampling
         break
      end
   end
end
waitbar(1);
close(w)
```

Measure forecast accuracy with continuous ranked probability score (CRPS)

The continuous ranked probability score (CRPS) is one of the most popular error scores for ensemble forecasts. It is strictly proper (Gneiting and Raftery, 2007), can be decomposed to understand reliability, resolution and uncertainty (Hersbach, 2000), and is in the units of measurement. Very usefully, CRPS collapses to the mean absolute error (MAE) for deterministic forecasts, providing a way to compare deterministic forecasts with ensemble forecasts.

In this case we use CRPS to compare the accuracy of our two forecasting systems, 'new' and 'old', at each lead time. CRPS values are calculated for every forecast with:

$$C(t) = \int_{-\infty}^{\infty} \left(F(t,x) - H(q_o(t) \leq x)\right)^2 dx \tag{51}$$

where $F(t,[\ ])$ is the cumulative distribution function (CDF) of the forecast ensemble at time t, $q_o(t)$ is the observation and H is the Heaviside step function. We calculate the average CRPS with:

$$\overline{C} = \frac{1}{T}\sum_{t=1}^{T} C(t) \tag{52}$$

To understand the uncertainty in this calculation, we bootstrap the averaging with 100 repeats, allowing us to present [0.05 0.95] confidence intervals.

In addition, we present the forecast accuracy as a skill score in comparison to a climatology forecast by:

$$C_{skill} = 1 - \frac{\overline{C_{fct}}}{\overline{C_{ref}}} \tag{53}$$

where $\overline{C_{fct}}$ is the mean CRPS for the forecast we are testing, and $\overline{C_{ref}}$ is the CRPS of our reference forecast, in this case climatology. We also estimate confidence in this calculation by bootstrapping with 100 repeats.

Calculate errors by lead time and bootstrap

```
% Define number of bootstrap repeats and confidence intervals
bsRepeats = 100;

% Preallocate memory
crpsNew = zeros(size(newFctQ.data,[1 4]))*nan;
crpsOld = crpsNew;
crpsClim = crpsNew;
meanCrpsNew = zeros(size(newFctQ.data,1),bsRepeats)*nan;
meanCrpsOld = meanCrpsNew;
meanCrpsClim = meanCrpsNew;

% Calculate errors
for i = 1:size(newFctQ.data,1)
   [~,crpsNew(i,:)] = crps(squeeze(newFctQ.data(i,newFctQ.stationId==stationId,:,:)),...
      verObs(i,:));
   [~,crpsOld(i,:)] = crps(squeeze(oldFctQ.data(i,newFctQ.stationId==stationId,:,:)),...
      verObs(i,:));
   [~,crpsClim(i,:)] = crps(squeeze(climFct(i,:,:)),verObs(i,:));
end

% Bootstrap average errors
for b = 1:bsRepeats
   randIntegers = randi(size(crpsNew,2),1,size(crpsNew,2));
   meanCrpsNew(:,b) = mean(crpsNew(:,randIntegers),2);
   meanCrpsOld(:,b) = mean(crpsOld(:,randIntegers),2);
   meanCrpsClim(:,b) = mean(crpsClim(:,randIntegers),2);
end

% Calculate skill against a climatological forecast
crpsSkillNew = 1-(meanCrpsNew./meanCrpsClim);
crpsSkillOld = 1-(meanCrpsOld./meanCrpsClim);
```

Plot forecast errors with lead time

```
% Specify confidence intervals for plotting
confIntPrctl = [5 50 95];
```

```
% Set up variables to plot
crpsNewPlot = prctile(meanCrpsNew,confIntPrctl,2);
crpsOldPlot = prctile(meanCrpsOld,confIntPrctl,2);

crpsSkillNewPlot = prctile(crpsSkillNew*100,confIntPrctl,2);
crpsSkillOldPlot = prctile(crpsSkillOld*100,confIntPrctl,2);

% Plot errors
figure(1); clf;
colOrder = get(gca,'colorOrder');

xvals = 1:size(meanCrpsNew,1);
ciplot(crpsNewPlot(:,1),crpsNewPlot(:,end),xvals,'color',colOrder(1,:));
hold on;
plot(xvals,crpsNewPlot(:,2),'color',colOrder(1,:))

ciplot(crpsOldPlot(:,1),crpsOldPlot(:,end),xvals,'color',colOrder(2,:));
hold on;
plot(xvals,crpsOldPlot(:,2),'color',colOrder(2,:))

% Add labels
xlabel(sprintf('lead time (%s)',leadTimeUnits))
ylabel(sprintf('CRPS (%s)',ncreadatt(obsNc,'q_obs','units')))
title('Forecast errors')

% Add annotation to help with interpretation of plot
xl = get(gca,'xlim');
yl = get(gca,'ylim');
os = 0.03;
pos = get(gca,'position');
annotation('arrow',[pos(1)+pos(3)+os*pos(3) pos(1)+pos(3)+os*pos(3)],...
   [pos(2)+pos(4) pos(2)]);
text(xl(end)+(os+0.01)*diff(xl),yl(1)+diff(yl)*0.5,sprintf('Better'));

% Add legend
ch = get(gca,'Children');
lh = legend(ch(end-1:-2:1),'New method','Old method');
lh.Location = 'northwest';
```

Figure 65 shows the forecast error against the lead time. We can see that CRPS increases with lead time for both new and old forecasting systems, which reduces accuracy.

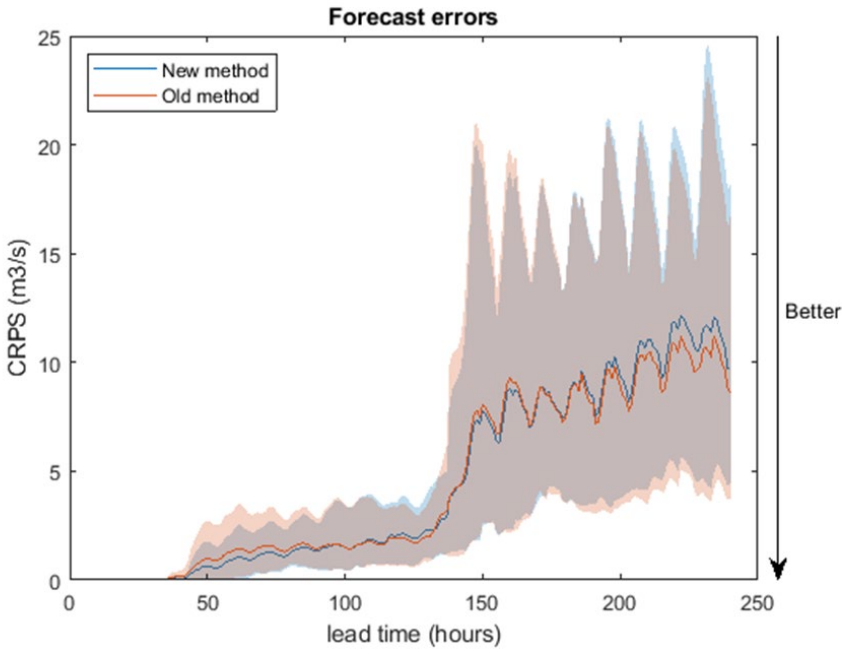


Figure 65. Mean CRPS of old (red) and new (blue) forecasts versus lead time

Plot skill with lead time

```
% Plot skill
figure(2); clf;
colOrder = get(gca,'colorOrder');

xvals = 1:size(meanCrpsNew,1);
ciplot(crpsSkillNewPlot(:,1),crpsSkillNewPlot(:,end),xvals,'color',colOrder(1,:));
hold on;
plot(xvals,crpsSkillNewPlot(:,2),'color',colOrder(1,:))

ciplot(crpsSkillOldPlot(:,1),crpsSkillOldPlot(:,end),xvals,'color',colOrder(2,:));
hold on;
plot(xvals,crpsSkillOldPlot(:,2),'color',colOrder(2,:))

% Add labels
xlabel(sprintf('lead time (%s)',leadTimeUnits))
ylabel('CRPS skill (%)')
title('Forecast skill compared to climatology')

% Add annotation to help with interpretation of plot
xl = get(gca,'xlim');
yl = get(gca,'ylim');
os = 0.03;
pos = get(gca,'position');
```

```
annotation('arrow',[pos(1)+pos(3)+os*pos(3) pos(1)+pos(3)+os*pos(3)],...
    [pos(2) pos(2)+pos(4)]);
text(xl(end)+(os+0.01)*diff(xl),yl(1)+diff(yl)*0.5,sprintf('Better'));

% Add legend
ch = get(gca,'children');
lh = legend(ch(end-1:-2:1),'New method','Old method');
lh.Location = 'northeast';
```

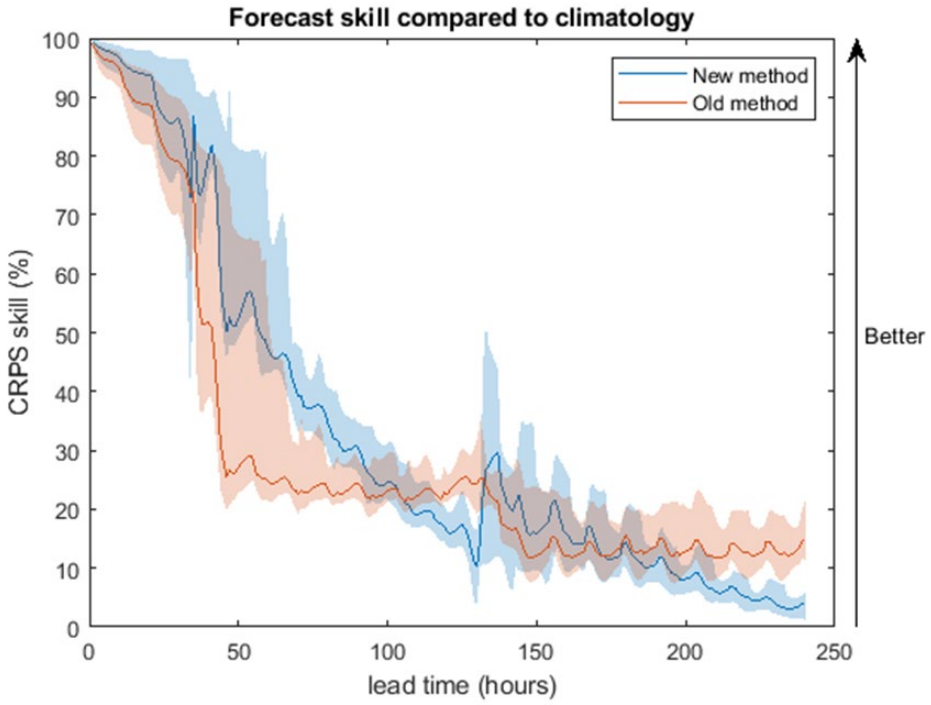


Figure 66. CRPSS in reference to climatology of old (red) and new (blue) forecasts versus lead time

In Figure 66 we can see that the forecasting systems are similarly accurate at most lead times, with the median error of the new method slightly larger than that of the old method at longer lead times (>~190 hours). CRPS increased with lead time, reflecting the declining influence of the autoregressive updating. The skill plot shows that the median skill of the new forecast is better than that of the old forecast for lead times of 0–~100 hours, while median skill of the old forecast is greater with lead times of >~190 hours. Both forecasting systems are positively skilful for the entire forecast period.

Assessing reliability with probability integral transform (PIT) uniformity in ephemeral streams

Reliability can be formally assessed by checking the uniformity of probability integral transform (PIT) values:

$$p(t) = \begin{cases} F\left(t, q_o(t)\right), & q_o(t) > 0 \\ U(0,1) \times F(t,0), & q_o(t) = 0 \end{cases} \tag{54}$$

where $F((t,[\ ])$ is the cumulative distribution function (CDF) of the forecast ensemble at time t, $q_o(t)$ is the observation and $U(0,1)$ is a random number drawn from the uniform distribution on

the range of (0, 1). The treatment at $q_o(t) = 0$ is required to ensure PIT values can be uniformly distributed even in the presence of zero values. When PIT values are generated in this way at $q_o(t)$, they are usually referred to as "pseudo PIT" values.

PIT values can be checked for uniformity by plotting them against a standard uniform variate (Figure 67), allowing us to assess the reliability of the ensemble. These plots are closely related to rank histograms/Talagrand diagrams (they are essentially slightly different representations of the same metric), but have the advantage of being suitable for assessing small numbers of forecasts. Histograms can be more difficult to construct and interpret with few data points. If forecasts are perfectly reliable, PIT values track the diagonal. Deviations from the diagonal can be interpreted as follows:

```
%% Generate and plot PIT interpretation plot
nFct = 5000;
nEns = 1000;
obs1 = randn(nFct,1)+100;
fct1 = randn(nEns,nFct)+100.75;
fct2 = randn(nEns,nFct)+99.25;
fct3 = randn(nEns,nFct)*0.5+100;
fct4 = randn(nEns,nFct)*2+100;

[PITv1,R1] = pit_ecdf(obs1,fct1);
[PITv2,R2] = pit_ecdf(obs1,fct2);
[PITv3,R3] = pit_ecdf(obs1,fct3);
[PITv4,R4] = pit_ecdf(obs1,fct4);

% Plot
figure(3); clf
col = get(gca,'colororder');
ms = 4;

plot(R1,PITv1,'o','markerfacecolor',col(1,:),'markersize',ms)
hold on;
plot(R2,PITv2,'o','markerfacecolor',col(2,:),'markersize',ms)
plot(R3,PITv3,'o','markerfacecolor',col(3,:),'markersize',ms)
plot(R4,PITv4,'o','markerfacecolor',col(4,:),'markersize',ms)
line([0 1], [0 1],'color','k')

set(gca,'plotboxaspectratio',[1 1 1],'layer','top')

% Add labels
ylabel('PIT')
xlabel('Std. Uniform Variate')
title('How to interpret PIT-uniform probability plots')

% Add legend
ch = get(gca,'children');
lh1 = legend(ch([5 4 1 3 2]),'Overprediction','Underprediction',...
   'Perfectly reliable','Ensemble too narrow','Ensemble too wide');
lh1.Location = "southoutside";
lh1.Orientation = "vertical";
lh1.NumColumns = 2;
```

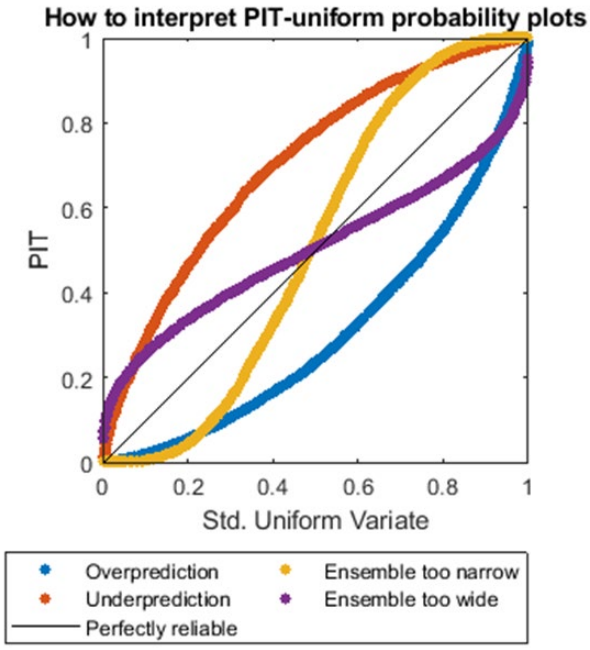


Figure 67. Interpretation of PIT plots

The significance of the deviation of PIT values from uniformity can be checked with Kolmogorov–Smirnoff significance bands. In this case, we set the significance to level to 5%. Note that the significance calculations assume independence between all forecast–observation pairs. Even though we are stratifying our PIT calculations by lead time, it is unlikely that all pairs are independent, because forecasts are issued twice per day. That is, streamflow data points 12 hours apart are likely to be correlated to some degree. For simplicity, we will ignore this problem here. It can be addressed by calculating PIT values only from independent values (for example, by only calculating PIT values from every 10th (or 20th) forecast) before significance tests are applied. This requires more than one month of forecasts to ensure robust metrics.

Calculate and plot PIT uniformity and significance

We will select four lead times – 1 , 48 , 96 and 216 hours – to see how reliability varies with lead time. In these plots we will identify pseudo-PIT values from conventionally calculated PIT to illustrate the importance of accounting for zero flow.

Calculate PIT

```
% Preallocate memory
pitValsNew = zeros(size(newFctQ.data,[1 4]))*nan;
pitValsOld = pitValsNew;

rValsNew = pitValsNew;
rValsOld = pitValsNew;

ksXnew = zeros([size(newFctQ.data,1) 2 2])*nan;
ksYnew = ksXnew;

ksXold = ksXnew;
ksYold = ksXnew;

pseudIndxNew = false(size(pitValsNew));
pseudIndxOld = pseudIndxNew;
```

```
% Calculate PIT
for i = 1:size(newFctQ.data,1)
   [pitValsNew(i,:),rValsNew(i,:),pseudIndxNew(i,:),ksXnew(i,:,:),ksYnew(i,:,:)] = ...
      pit_ecdf(verObs(i,:),squeeze(newFctQ.data(i,newFctQ.stationId==stationId,:,:)));
   [pitValsOld(i,:),rValsOld(i,:),pseudIndxOld(i,:),ksXold(i,:,:),ksYold(i,:,:)] = ...
      pit_ecdf(verObs(i,:),squeeze(oldFctQ.data(i,newFctQ.stationId==stationId,:,:)));
end
```

Plot PIT uniformity

```
plotLead = [1 48 96 216];
figure(4); clf;
colOrder = get(gca,'colorOrder');
sp = nan(size(plotLead));
for p = 1:length(plotLead)
   tIndx = plotLead(p);
   p1 = p; if p>2; p1 = p+1; end
   sp(p) = subplot(2,ceil(length(plotLead)/2)+1,p1);
   plot(rValsNew(tIndx,~pseudIndxNew(i,:)), pitValsNew(tIndx,~pseudIndxNew(i,:)),...
      '.','color',colOrder(1,:));
   hold on;
   plot(rValsNew(tIndx,pseudIndxNew(i,:)),pitValsNew(tIndx,pseudIndxNew(i,:)),...
      '+','color',colOrder(1,:));
   plot(rValsOld(tIndx,~pseudIndxOld(i,:)), pitValsOld(tIndx,~pseudIndxOld(i,:)),...
      '.','color',colOrder(2,:));
   plot(rValsOld(tIndx,pseudIndxOld(i,:)),pitValsOld(tIndx,pseudIndxOld(i,:)),...
      '+','color',colOrder(2,:));
   plot(squeeze(ksXnew(i,:,:)),squeeze(ksYnew(i,:,:)),':k');
   line; xlim([0 1]); ylim([0 1]);
   set(gca,'plotboxaspectratio',[1 1 1])
   title(sprintf('lead %d h',plotLead(p)))
   if p == length(plotLead)
      ch = get(gca,'children');
      lh = legend(ch([end:-1:end-3 2]),'New PIT','New pseudo-PIT','Old PIT',...
         'Old pseudo-PIT','KS significance','orientation','vertical');
      pos = get(sp(end),'InnerPosition');
      os = 1.1;
      lh.Position = [pos(1)+pos(3)*os pos(2)+(os-1)*pos(4) ...
         lh.Position(3) lh.Position(4)];
   end
   if mod(p,2)==1
      ylabel('PIT')
   end
   if p>2
      xlabel('Std. Unif. Var.')
   end
end
```

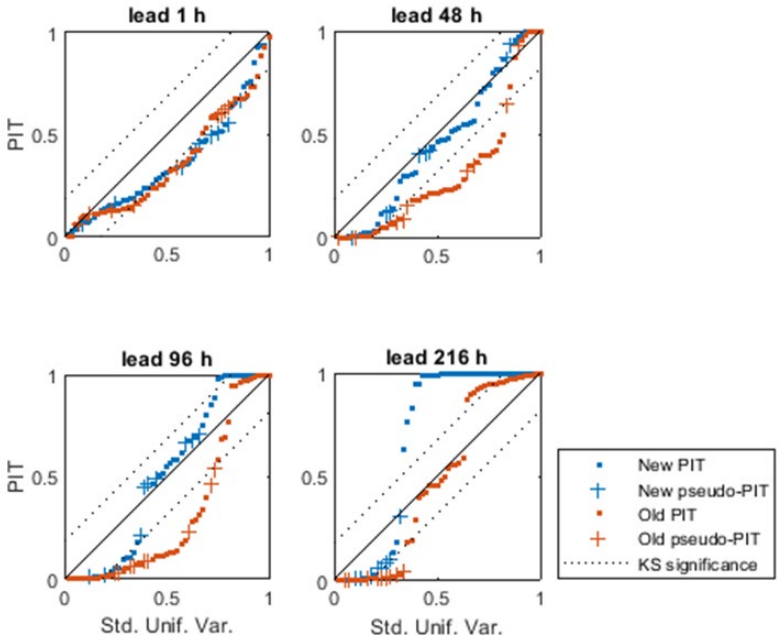


Figure 68. PIT plots of old (red) and new (blue) forecasts at various lead times with Kolmogorov–Smirnoff significance bands

We can see from the plots in Figure 68 that the forecasting systems produce different levels of reliability, and also that reliability varies substantially with lead time. Forecasts from the new system are largely reliable at lead times of 48 hours, but are positively biased at 1 hour, and become increasingly overconfident at longer lead times, as shown by the plots of lead times of 96 and 216 hours. Forecasts from the old system tend to be positively biased at lead times of 1, 48 and 96 hours. The old forecasting system achieves its best reliability at lead 216 hours but is slightly overconfident.

PIT uniformity summary statistics

It is often convenient to summarize the uniformity of PIT values (for example, to summarize reliability at different lead times), as we will do below. A popular statistic to do this is Renard et al.'s (2010) α-index. This calculates the deviation of PIT values from the ideal theoretical uniform distribution, given by:

$$\alpha = 1 - \frac{2}{T}\sum_{t=1}^{T}|p(t) - p_U(t)| \tag{55}$$

Where $P_U(t)$ is the theoretical value corresponding to $p(t)$ (that is, the value drawn from the standard uniform variate in Figure 68). α ranges from 1 (perfectly reliable) to ∞. As with CRPS, we will bootstrap this calculation.

The α-index effectively calculates deviation from the diagonal in Figure 68. This is somewhat reductive, as it does not give any information on whether poor reliability is caused by bias or incorrectly specified ensemble spread. An alternative pair of indices is the β-score and β-bias by Keller and Hense (2011). This diagnostic fits parametric distributions to rank histograms to determine the degree to which deficiencies in ensemble spread (β-score) or bias (β-bias) contribute to poor reliability. This method is more complex to describe, and we refer the reader to the original paper for details. The β-score ranges from −∞ to ∞. Values near zero indicate more reliable forecasts, with values >0 indicating under-confidence (ensemble is too wide) and values <0 indicating overconfidence (ensemble is too narrow). β-bias also ranges from −∞ to ∞. Values near zero indicate more unbiased forecasts, with values >0 indicating positive biases and values <0 indicating negative biases.

Calculate *α*-index with bootstrapping

```
% Set number of bootstrap repeats
bsRepeats = 100;

% Preallocate memory
alphaNew = zeros(size(pitValsNew,1),bsRepeats);
alphaOld = alphaNew;

% Bootstrap average errors
for i = 1:size(pitValsNew,1)
   for b = 1:bsRepeats
      randIntegers = randi(size(pitValsNew,2),1,size(pitValsNew,2));

      na = length(pitValsNew(i,randIntegers));
      alphaPrime = sum(abs(pitValsNew(i,randIntegers) - rValsNew(i,randIntegers)))/na;
      alphaNew(i,b) = 1-2*alphaPrime;

      na = length(pitValsNew(i,randIntegers));
      alphaPrime = sum(abs(pitValsOld(i,randIntegers) - rValsOld(i,randIntegers)))/na;
      alphaOld(i,b) = 1-2*alphaPrime;
   end
end
```

Plot *α*-index with lead time

```
% Specify confidence intervals for plotting
confIntPrctl = [5 50 95];

% Set up variables to plot
alphaNewPlot = prctile(alphaNew,confIntPrctl,2);
alphaOldPlot = prctile(alphaOld,confIntPrctl,2);

% Plot
figure(5); clf;
colOrder = get(gca,'colorOrder');

xvals = 1:size(pitValsNew,1);
ciplot(alphaNewPlot(:,1),alphaNewPlot(:,end),xvals,'color',colOrder(1,:));
hold on;
plot(xvals,alphaNewPlot(:,2),'color',colOrder(1,:))

ciplot(alphaOldPlot(:,1),alphaOldPlot(:,end),xvals,'color',colOrder(2,:));
hold on;
plot(xvals,alphaOldPlot(:,2),'color',colOrder(2,:))

% Add labels
xlabel(sprintf('lead time (%s)',leadTimeUnits))
yl = ylabel('$$\alpha$$-index');
yl.Interpreter = 'latex';
title('Forecast reliability')
```

```
% Add annotation to help with interpretation of plot
xl = get(gca,'xlim');
yl = get(gca,'ylim');
os = 0.03;
pos = get(gca,'position');
annotation('arrow',[pos(1)+pos(3)+os*pos(3) pos(1)+pos(3)+os*pos(3)],...
    [pos(2) pos(2)+pos(4)]);
text(xl(end)+(os+0.01)*diff(xl),yl(1)+diff(yl)*0.5,sprintf('Better'));

% Add legend
ch = get(gca,'children');
lh = legend(ch(end-1:-2:1),'New method','Old method');
lh.Location = 'northeast';
```

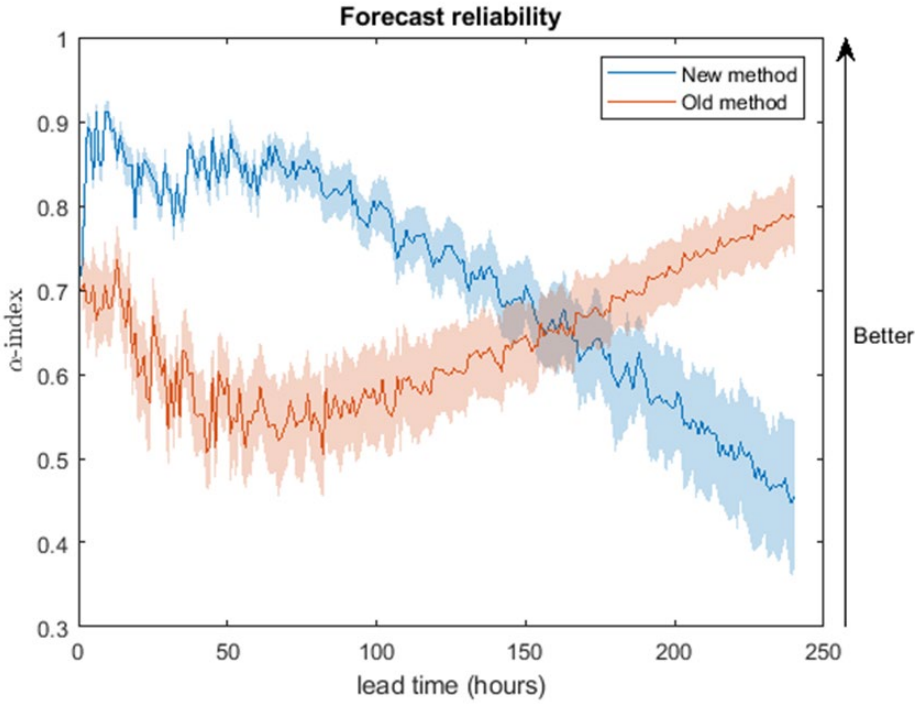


Figure 69. *α*-index of old (red) and new (blue) forecasts versus lead time

As we have seen with the PIT-uniform probability plots, forecast reliability varies with lead time. For the new forecasting system, the forecasts are quite reliable for the first ~75 hours, after which reliability declines (Figure 69). For the old forecasting system, positive biases cause poorer reliability than with the new system until lead times of ~150 hours. After this point, the new forecasts become increasingly overconfident, leading to less reliable forecasts than in the old system.

Calculate *β*-score and *β*-bias

The *β*-score and *β*-bias are calculated on rank histograms. We use PIT values as ranks, but we must decide how many "bins" (bars in the histogram) these ranks will be divided into. We choose 8 bins, as this would allow 7 forecasts to be grouped in each bin if the histograms are totally uniform (as we have 56 forecasts). Calculating the two *β* metrics involves an optimization to fit the parametric *β* distribution to the shape of the rank histogram. This can be a somewhat computationally intensive, and bootstrapping this estimation with 100 repeats can take a few minutes.

```
% Set number of bins
nBin = 8;

% Set number of bootstrap repeats
bsRepeats = 100;

% Preallocate memory
betaNew = zeros(2,size(pitValsNew,1),bsRepeats)*nan;
betaOld = betaNew;

w = waitbar(0,'fitting beta and bootstrapping...','Name','beta score/beta bias');
for i = 1:size(pitValsNew,1)
   for b = 1:bsRepeats
      waitbar(((i-1)*bsRepeats+b)/(size(pitValsNew,1)*bsRepeats));
      randIntegers = randi(size(pitValsNew,2),1,size(pitValsNew,2));

      [~,~,dumBeta] = compute_betascore(pitValsNew(i,randIntegers),nBin,true);
      betaNew(1,i,b) = dumBeta.beta_score;
      betaNew(2,i,b) = dumBeta.beta_bias;

      [~,~,dumBeta] = compute_betascore(pitValsOld(i,randIntegers),nBin,true);
      betaOld(1,i,b) = dumBeta.beta_score;
      betaOld(2,i,b) = dumBeta.beta_bias;
   end
end
waitbar(1);
close(w);
```

Plot β-score and β-bias

```
% Specify confidence intervals for plotting
confIntPrctl = [5 50 95];
yl = [-2 2]; %ylimits

% Set up variables to plot
betaScoreNewPlot = prctile(squeeze(betaNew(1,:,:)),confIntPrctl,2);
betaScoreOldPlot = prctile(squeeze(betaOld(1,:,:)),confIntPrctl,2);

betaBiasNewPlot = prctile(squeeze(betaNew(2,:,:)),confIntPrctl,2);
betaBiasOldPlot = prctile(squeeze(betaOld(2,:,:)),confIntPrctl,2);

% Plot
figure(6); clf;
colOrder = get(gca,'colorOrder');

subplot(2,1,1)
xvals = 1:size(betaScoreNewPlot,1);
ciplot(betaScoreNewPlot(:,1),betaScoreNewPlot(:,end),xvals,'color',colOrder(1,:));
hold on;
plot(xvals,betaScoreNewPlot(:,2),'color',colOrder(1,:))
```

```
ciplot(betaScoreOldPlot(:,1),betaScoreOldPlot(:,end),xvals,'color',colOrder(2,:));
hold on;
plot(xvals,betaScoreOldPlot(:,2),'color',colOrder(2,:))
ylim(yl)
xl = get(gca,'xlim');
line(xl,[0 0],'color','k')

% Add labels
yl = ylabel('$$\beta$$-score');
yl.Interpreter = 'latex';
title('Appropriateness of ensemble spread')

% Add annotation to help with interpretation of plot
xl = get(gca,'xlim');
yl = get(gca,'ylim');
os = 0.03;
pos = get(gca,'position');
annotation('arrow',[pos(1)+pos(3)+os*pos(3) pos(1)+pos(3)+os*pos(3)],...
   [pos(2)+pos(4)/2 pos(2)+pos(4)]);
text(xl(end)+(os+0.01)*diff(xl),yl(1)+diff(yl)*0.75,sprintf('under-\nconfident'));

annotation('arrow',[pos(1)+pos(3)+os*pos(3) pos(1)+pos(3)+os*pos(3)],...
   [pos(2)+pos(4)/2 pos(2)]);
text(xl(end)+(os+0.01)*diff(xl),yl(1)+diff(yl)*0.25,sprintf('over-\nconfident'));

% Add legend
ch = get(gca,'children');
lh = legend(ch([end-1 end-3]),'New method','Old method');
lh.Location = 'northeast';

subplot(2,1,2)
ciplot(betaBiasNewPlot(:,1),betaBiasNewPlot(:,end),xvals,'color',colOrder(1,:));
hold on;
plot(xvals,betaBiasNewPlot(:,2),'color',colOrder(1,:))

ciplot(betaBiasOldPlot(:,1),betaBiasOldPlot(:,end),xvals,'color',colOrder(2,:));
hold on;
plot(xvals,betaBiasOldPlot(:,2),'color',colOrder(2,:))
ylim(yl)
xl = get(gca,'xlim');
line(xl,[0 0],'color','k')

% Add labels
xlabel(sprintf('lead time (%s)',leadTimeUnits))
yl = ylabel('$$\beta$$-bias');
yl.Interpreter = 'latex';
title('Bias of ensemble spread')

% Add annotation to help with interpretation of plot
xl = get(gca,'xlim');
```

```
yl = get(gca,'ylim');
os = 0.03;
pos = get(gca,'position');
annotation('arrow',[pos(1)+pos(3)+os*pos(3) pos(1)+pos(3)+os*pos(3)],...
   [pos(2)+pos(4)/2 pos(2)+pos(4)]);
text(xl(end)+(os+0.01)*diff(xl),yl(1)+diff(yl)*0.75,sprintf('+ve bias'));

annotation('arrow',[pos(1)+pos(3)+os*pos(3) pos(1)+pos(3)+os*pos(3)],...
   [pos(2)+pos(4)/2 pos(2)]);
text(xl(end)+(os+0.01)*diff(xl),yl(1)+diff(yl)*0.25,sprintf('-ve bias'));
```

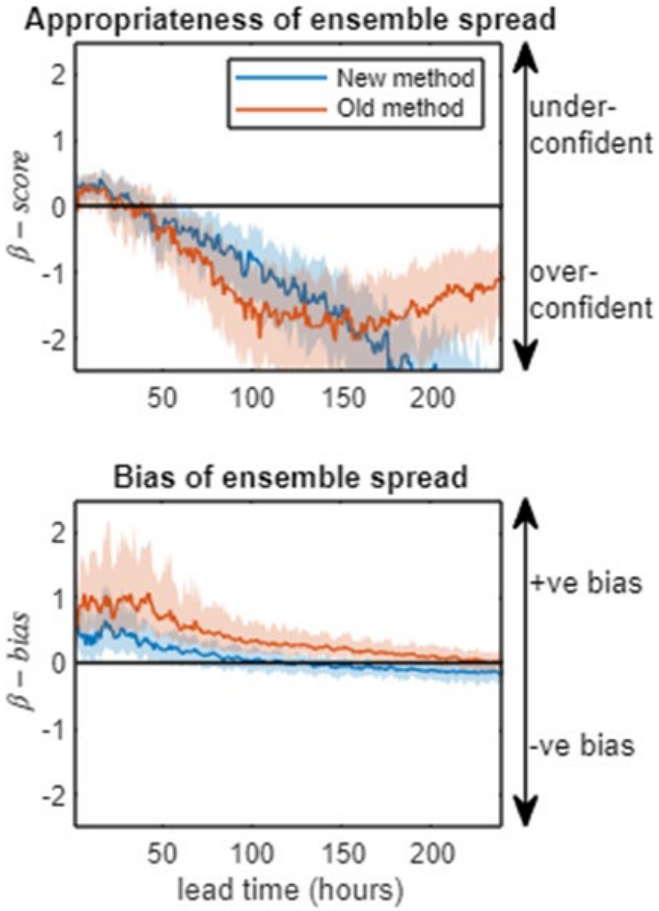


Figure 70. β-score (top) and β-bias (bottom) of old (red) and new (blue) forecasts versus lead time

We gain additional insights from the plot in Figure 70, consistent with our inspection of the PIT-uniform probability plots above. We can see that both methods exhibit overconfidence for lead times >~50 hours. Overconfidence is the major fault with the new method, particularly at lead times >150 hours, where the median β-score drops below the axis of the chart. β-bias values for the new forecasts are close to zero at longer lead times. The old method tends to suffer more from positive bias, particularly at earlier lead times.

Sharpness

Finally, we wish our forecasts to be as sharp as possible, while still being reliable. Sharpness is often expressed as the average width of prediction intervals (AWPI):

$$AWPI = \frac{1}{T}\sum_{t=1}^{T}\left(F^{-1}\left(t, 1-\left(\frac{1-\Delta}{2}\right)\right) - F^{-1}\left(t, \left(\frac{1-\Delta}{2}\right)\right)\right) \quad (56)$$

where $F^{-1}(t,[\ ])$ is the inverse cumulative distribution function (CDF) of the forecast ensemble at time t, and Δ is the confidence interval. Lower values are better. Choosing a single interval is somewhat reductive, as AWPI can be sensitive to the interval chosen. We therefore choose three intervals: $\Delta = 50\%$ (that is, the interquartile range), $\Delta = 80\%$ and $\Delta = 90\%$. As with the a-index and CRPS, it is possible to bootstrap the calculation to offer confidence intervals in the score.

We wish our forecasts to be at least as sharp as climatology and accordingly calculate an AWPI ratio as a skill score:

$$AWPI\ Ratio = 1 - \frac{AWPI_{Fct}}{AWPI_{Ref}} \quad (57)$$

where $AWPI_{Fct}$ is the AWPI of the forecast being assessed, and $AWPI_{Ref}$ is the AWPI of our climatology reference forecast. The AWPI ratio ranges from 1 (perfectly sharp) to −∞.

Calculate and bootstrap average width of prediction intervals (AWPI)

```
% Define number of bootstrap repeats and confidence intervals
bsRepeats = 100;
predInt = [50 80 90];

% Preallocate memory
wpiNew = zeros(size(newFctQ.data,1),size(newFctQ.data,4),length(predInt))*nan;
wpiOld = wpiNew;
wpiClim = wpiNew;

awpiNew = zeros(size(newFctQ.data,1),bsRepeats,length(predInt))*nan;
awpiOld = awpiNew;
awpiClim = awpiNew;

% Calculate AWPI
for p = 1:length(predInt)
   piPrctl = [100-(100-predInt(p))/2 (100-predInt(p))/2];
   wpiNew(:,:,p) = squeeze(prctile(newFctQ.data(:,newFctQ.stationId==stationId,:,:),...
      piPrctl(1),3)-prctile(newFctQ.data(:,newFctQ.stationId==stationId,:,:),...
      piPrctl(2),3));
   wpiOld(:,:,p) = squeeze(prctile(oldFctQ.data(:,oldFctQ.stationId==stationId,:,:),...
      piPrctl(1),3)-prctile(oldFctQ.data(:,oldFctQ.stationId==stationId,:,:),...
      piPrctl(2),3));
   wpiClim(:,:,p) = squeeze(prctile(climFct,piPrctl(1),2)-prctile(climFct,...
      piPrctl(2),2));
end

for b = 1:bsRepeats
   randIntegers = randi(size(wpiNew,2),1,size(wpiNew,2));
   awpiNew(:,b,:) = mean(wpiNew(:,randIntegers,:),2);
   awpiOld(:,b,:) = mean(wpiOld(:,randIntegers,:),2);
   awpiClim(:,b,:) = mean(wpiClim(:,randIntegers,:),2);
end
```

```
% Calculate skill against a climatolology forecast
awpiRatioNew = 1-(awpiNew./awpiClim);
awpiRatioOld = 1-(awpiOld./awpiClim);
```

Plot AWPI with lead time

```
% Specify confidence intervals for plotting
confIntPrctl = [5 50 95];

% Set up variables to plot
awpiNewPlot = squeeze(prctile(awpiNew,confIntPrctl,2));
awpiOldPlot = squeeze(prctile(awpiOld,confIntPrctl,2));

% Plot AWPI
figure(7); clf;
colOrder = get(gca,'colorOrder');

xvals = 1:size(awpiNew,1);

for p = 1:length(predInt)
   subplot(length(predInt),1,p)

   ciplot(awpiNewPlot(:,1,p),awpiNewPlot(:,end,p),xvals,'color',colOrder(1,:));
   hold on;
   plot(xvals,awpiNewPlot(:,2,p),'color',colOrder(1,:))

ciplot(awpiOldPlot(:,1,p),awpiOldPlot(:,end,p),xvals,'color',colOrder(2,:));
   hold on;
   plot(xvals,awpiOldPlot(:,2,p),'color',colOrder(2,:))

   % Add labels
   if p == length(predInt)
      xlabel(sprintf('lead time (%s)',leadTimeUnits))
   end
   ylabel(sprintf('AWPI (%s)',ncreadatt(obsNc,'q_obs','units')))
   title(sprintf('AWPI of %d%% interval',predInt(p)))

   % Add annotation to help with interpretation of plot
   xl = get(gca,'xlim');
   yl = get(gca,'ylim');
   os = 0.03;
   pos = get(gca,'position');
   annotation('arrow',[pos(1)+pos(3)+os*pos(3) pos(1)+pos(3)+os*pos(3)],...
      [pos(2)+pos(4) pos(2)]);
   text(xl(end)+(os+0.01)*diff(xl),yl(1)+diff(yl)*0.5,sprintf('Better'));

   if p ==1
      % Add legend
      ch = get(gca,'Children');
      lh = legend(ch(end-1:-2:1),'New method','Old method');
      lh.Location = 'northwest';
```

```
    end
end
```

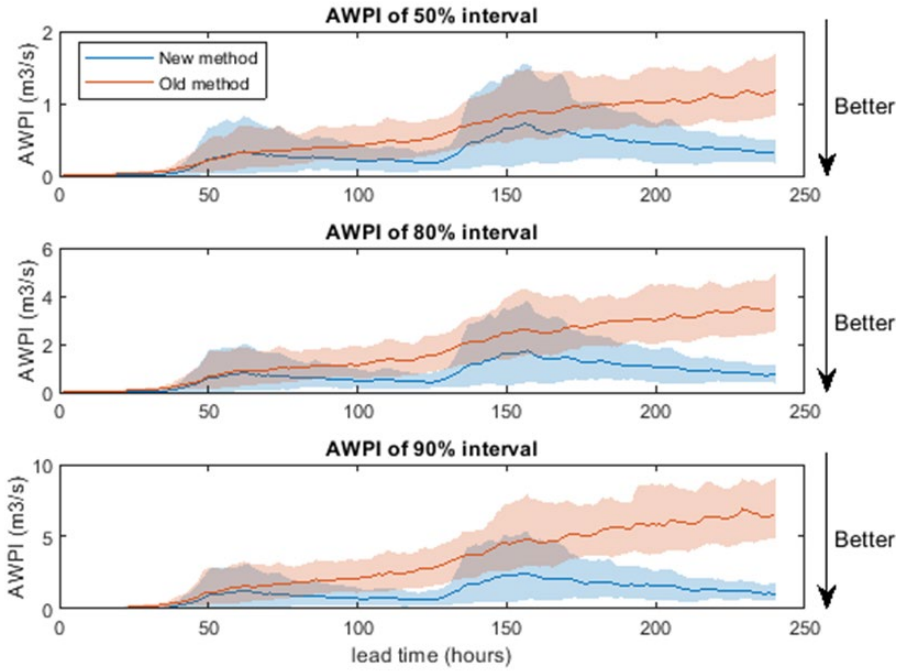


Figure 71. AWPI of old (red) and new (blue) forecasts versus lead time at 50% (top), 80% (middle) and 90% (bottom) confidence intervals

In Figure 71 we can see that the new forecasting system is generally sharper than the old forecasting system, particularly at longer lead times. This holds for all three prediction intervals tested. However, this comes at the expense of reliability at longer lead times (>~150 hours). Sharpness without reliability is highly undesirable: if it were not, we would reward deterministic forecasts.

Plot AWPI ratio

```
% Specify confidence intervals for plotting
confIntPrctl = [5 50 95];

% Set up variables to plot
awpiRatioNewPlot = squeeze(prctile(awpiRatioNew,confIntPrctl,2))*100;
awpiRatioOldPlot = squeeze(prctile(awpiRatioOld,confIntPrctl,2))*100;

% Plot AWPI ratio
figure(8); clf;
colOrder = get(gca,'colorOrder');

xvals = 1:size(meanCrpsNew,1);

for p = 1:length(predInt)
    subplot(length(predInt),1,p)
    ciplot(awpiRatioNewPlot(:,1,p),awpiRatioNewPlot(:,end,p),xvals,...
        'color',colOrder(1,:));
    hold on;
```

```
    plot(xvals,awpiRatioNewPlot(:,2,p),'color',colOrder(1,:))

    ciplot(awpiRatioOldPlot(:,1,p),awpiRatioOldPlot(:,end,p),xvals,...
        'color',colOrder(2,:));
    hold on;
    plot(xvals,awpiRatioOldPlot(:,2,p),'color',colOrder(2,:))
    ylim([-100 100])
    xl = get(gca,'xlim');
    line(xl,[0 0],'color','k')

    % Add labels
    if p == length(predInt)
        xlabel(sprintf('lead time (%s)',leadTimeUnits))
    end
    ylabel('AWPI ratio (%)')
    title(sprintf('AWPI Ratio for %d%% interval',predInt(p)))

    % Add annotation to help with interpretation of plot
    xl = get(gca,'xlim');
    yl = get(gca,'ylim');
    os = 0.03;
    pos = get(gca,'position');
    annotation('arrow',[pos(1)+pos(3)+os*pos(3) pos(1)+pos(3)+os*pos(3)],...
        [pos(2)+pos(4)/2 pos(2)+pos(4)]);
    text(xl(end)+(os+0.01)*diff(xl),yl(1)+diff(yl)*0.75,sprintf('Sharp'));

    annotation('arrow',[pos(1)+pos(3)+os*pos(3) pos(1)+pos(3)+os*pos(3)],...
        [pos(2)+pos(4)/2 pos(2)]);
    text(xl(end)+(os+0.01)*diff(xl),yl(1)+diff(yl)*0.25,sprintf('Not\nsharp'));

    % Add legend
    if p ==1
        ch = get(gca,'children');
        lh = legend(ch([end-1 end-3]),'New method','Old method');
        lh.Location = 'southwest';
    end
end
```

The new method produces forecasts that are largely sharper than climatology for all lead times (Figure 72). The old method becomes less sharp than climatology after lead times of ~140–170 hours, depending on choice of prediction interval. Once again, however, the sharpness of the new forecasting system at longer lead times comes at the expense of reliability.

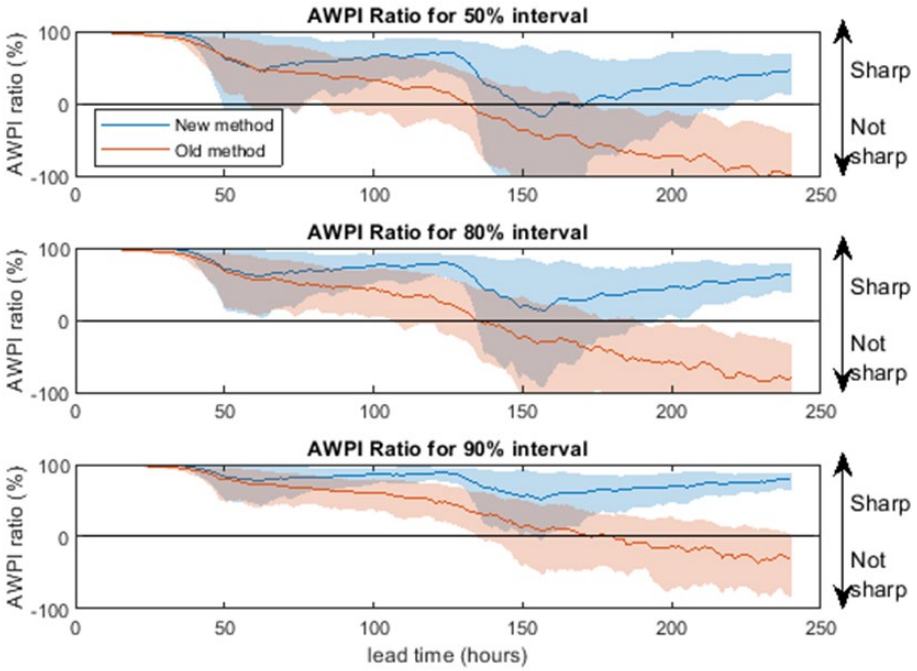


Figure 72. Same as Figure 71 but for AWPI ratio

APPENDIX B. DISTRIBUTIONS-ORIENTED APPROACH

Murphy and Winkler (1987) introduced a general verification framework referred to as the distributions-oriented (DO) approach in which the forecast and the verifying observation are treated as random variables and each forecast–observation pair is assumed to be independent of all other pairs and identically distributed (IID). The IID assumption means that, for verification, one only needs to describe a single joint relationship between the two random variables representing the forecast and the observation. The purpose of this appendix is to provide mathematical descriptions of the above relationship and the attendant concepts and definitions using basic probability theory for engineers (Drake, 1967; Benjamin and Cornell, 1970). Hence, unlike in the main text, estimation of statistical quantities or statistical inferences are not of concern in this appendix.

B.1 Joint, conditional and marginal probability distributions

Let X and Y be random variables representing the forecast and the observation of the variable of interest, and x and y denote the experimental values, or outcomes, that X and Y may take on, respectively. The probabilistic relationship between the forecast, X, and the observation, Y, is described wholly by their joint, or bivariate, probability density function (PDF), $f_{X,Y}(x,y)$, for continuous random variables, and probability mass function (PMF), $p_{X,Y}(x,y)$, for discrete random variables. The variables of interest for hydrological verification may be continuous, discrete or of mixed type. To aid intuitive understanding, X and Y are treated in this appendix as discrete random variables with little loss of generality. The only notational difference is that, for continuous variables, integration replaces summation.

The joint cumulative distribution function (CDF) of X and Y, $F_{X\leq,Y\leq}(x,y)$ is the cumulative sum of the probability masses in the joint PMF:

$$F_{X\leq,Y\leq}(x,y) = \Pr[X \leq x, Y \leq y] = \Sigma_{y_o \leq y} \Sigma_{x_o \leq x}\, p_{X,Y}(x_o, y_o) \tag{58}$$

where Pr[ ] denotes the probability that the event bracketed may occur. Note in Equation 58 that the summation is only over the experimental values less than or equal to the upper limits, x and y, and x_o and y_o are dummy variables. The joint PMF may be written as a product of marginal and conditional PMFs (Drake, 1967; Benjamin and Cornell, 1970):

$$p_{X,Y}(x,y) = p_{Y|X}(y|x)\, p_X(x) = p_{X|Y}(x|y)\, p_Y(y) \tag{59}$$

The first equality in Equation 59, which factors out the conditional PMF of observation given forecast, $p_{Y|X}(y|x)$, is referred to as the calibration–refinement (CR) factorization in the verification literature. The second equality in Equation 59, which factors out the conditional PMF of forecast given observation, $p_{X|Y}(x|y)$, is referred to as the likelihood–base rate (LBR) factorization. The marginal PMF of the forecast, $p_X(x)$, is referred to as the predictive or refinement distribution (Wilks, 2011). The marginal PMF of the observation, $p_Y(y)$, is referred to as the base rate or uncertainty distribution. When the random variables are continuous, Equation 59 still holds but with the PMFs replaced with the respective PDFs. In the rest of this appendix, the term distribution is used to refer to both PMF and PDF for simplicity.

Figure 73 illustrates the joint, marginal and scaled conditional distributions of discrete forecast and observation. The joint distribution $p_{X,Y}(x,y)$ is represented in the large box by the image plot for which, the darker the shade is, the larger the probability is. Each small square in the image plot represents the probability that the discrete random variables X and Y may take on the experimental value, or outcome, x and y, respectively. The histogram at the top of the figure represents the marginal distribution of the forecast, $p_X(x)$. The histogram to the right of the image plot represents the marginal distribution of the observation, $p_Y(y)$.

The vertical column encased in thick solid lines within the image plot represents the conditional distribution $p_{Y|X}(y|x_o)$, where x_o is some specific value of x, scaled by a constant $p_X(x_o)$ in accordance with Equation 59. Hence, each vertical strip in the image plot depicts the relative likelihood of different outcomes being observed given the specific forecast event, x_o. The row encased in thick solid lines represents the conditional distribution $p_{X|Y}(x|y_o)$, where y_o is some specific value of y, scaled by a constant $p_Y(y_o)$ in accordance with Equation 59. Hence, each horizontal strip in the image plot depicts the relative likelihood of different outcomes being forecast given the specific observed event, y_o.

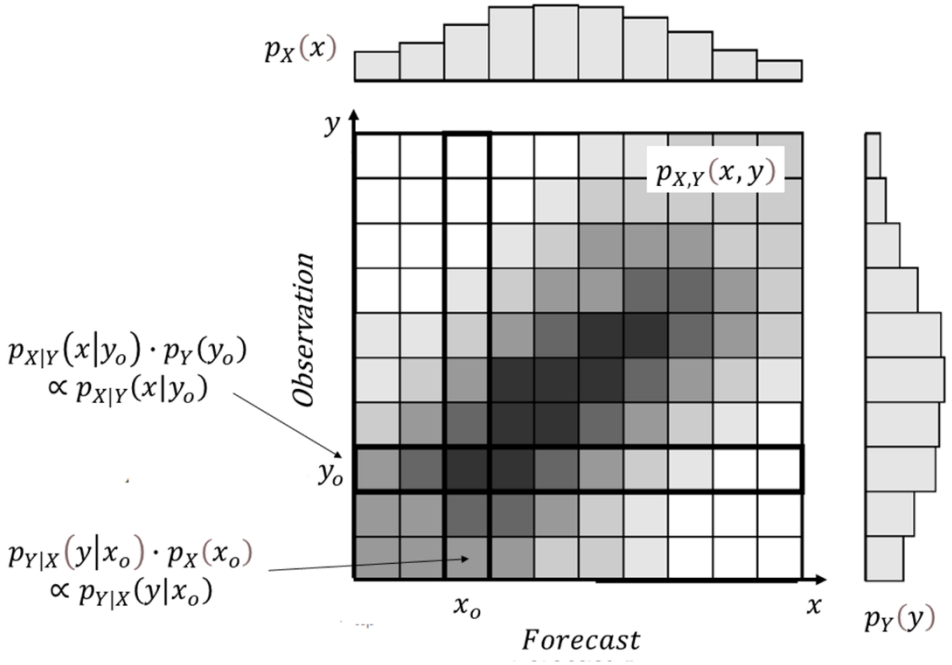


Figure 73. Conceptual illustration, for discrete random variables *X* and *Y*, of the joint distribution $p_{X,Y}(x,y)$, its marginal distributions, $p_X(x)$ and $p_Y(y)$, and the scaled conditional distributions, $p_{X|Y}(x|y_o)$ and $p_{Y|X}(y|x_o)$.

Source: Adapted from Bradley et al. (2019)

B.2 Expectations and moments

Various low-order mathematical moments of the random variables X and Y are used to summarize the key attributes of their joint distribution. Following is a partial list of those referred to in this publication with their definitions.

The first moment, which is also referred to as first-order moment, expectation, expected value or mean, of the forecast, X, is given by:

$$\mu_X = E[X] = \Sigma_x \, x \, p_X(x) \tag{60}$$

where the summation is for all possible experimental values of X (that is, all events that the forecast system can predict). Similarly, the first moment of the observation, Y, is given by:

$$\mu_Y = E[Y] = \Sigma_y \, y \, p_Y(y) \tag{61}$$

where the summation is for all possible outcomes of the observation, Y, regardless of whether the forecast system can predict or not. The conditional expectation, or conditional mean, of the forecast given observation, Y, is given by:

$$\mu_{X|Y}=E[X|Y]=\Sigma_x \, x \, p_{X|Y}(x|y) \tag{62}$$

Similarly, the conditional expectation of the observation, *Y*, given forecast, *X*, is given by:

$$\mu_{Y|X}=E[Y|X]=\Sigma_Y\ y\ p_{Y|X}(y|x) \quad (63)$$

Note in Equations 62 and 63 that $\mu_{X|Y}$ and $\mu_{Y|X}$ are conditioned on random variables *Y* and *X*, respectively. Being functions of *Y* and *X*, $\mu_{X|Y}$ and $\mu_{Y|X}$ are random variables themselves.

The centred second moment, or variance, of *X* is given by:

$$\sigma_X^2 = E[(X-\mu_X)^2] = E[X^2] - \mu_X^2 = \Sigma_x(x-\mu_X)^2 p_X(x) \quad (64)$$

Similarly, variance of *Y* is given by:

$$\sigma_Y^2 = E[(Y-\mu_Y)^2] = E[Y^2] - \mu_Y^2 = \Sigma_y(y-\mu_Y)^2 p_Y(y) \quad (65)$$

The centred cross-moment, or covariance, between *X* and *Y* is given by:

$$Cov[X,Y] = E[(X-\mu_X)(Y-\mu_Y)] = \Sigma_x\Sigma_y(x-\mu_X)(y-\mu_Y)\ p_{X,Y}(x,y) \quad (66)$$

Correlation, or Pearson correlation, between *X* and *Y* is given by:

$$\rho_{X,Y} = Cov[X,Y]/(\sigma_X\sigma_Y) \quad (67)$$

where σ_X and σ_Y denote the standard deviations of *X* and *Y*, respectively. It is important to recognize that the above moments completely prescribe the bivariate normal and lognormal distributions and other commonly used two-parameter distributions such as bivariate gamma (Nagao and Kadoya, 1970; Iliopoulos et al., 2005) and Weibull (Lu and Bhattacharyya, 1990; Johnson et al., 1999), which are widely used in hydrological applications. Hence, if the forecast and observation are indeed a jointly IID random process, follow one of the parametric bivariate distributions above, and have sufficiently large samples (that is, experimental outcomes), verification should be a straightforward endeavour.

B.3 Forecast attributes

The mathematical expressions for the key attributes are given below for the single-valued forecast versus the verifying observation using the mean squared error (MSE) as the measure of accuracy. The same expressions also apply to probability forecasts with the Brier score (BS) as the accuracy measure. The attributes are grouped according to three different decompositions of the MSE.

Mean bias, second-order bias and association

The first decomposition expresses the MSE in terms of the contributions from mean bias, bias in standard deviation and strength of linear association (Murphy and Winkler, 1987; Nelson et al., 2010):

$$MSE = E[(X-Y)]^2 = (\mu_X-\mu_Y)^2 + (\sigma_X-\sigma_Y)^2 + 2\sigma_X\sigma_Y\big(1-\rho_{X,Y}\big) \quad (68)$$

In Equation 68, the first term is the mean bias, or mean error (ME), squared, the second term is the bias in standard deviation squared and the third term represents the strength of association. If the marginal distributions of *X* and *Y* are identical, the first two terms vanish. The third term shows how linear correlation contributes to second-order accuracy. If the forecast is perfectly correlated, only the mean bias and the bias in standard deviation determine accuracy. As correlation decreases, the third term, and hence the MSE, increase. If the forecast is negatively correlated with the observation, the third term becomes larger than when the forecast has no correlation with the observation.

With Equation 68, one is now in a position to contrast MSE with the Kling–Gupta efficiency (KGE) (Gupta et al., 2009). The KGE is defined as:

$$KGE = 1 - \sqrt{\left(\rho_{X,Y} - 1\right)^2 + \left(\frac{\mu_X}{\mu_Y} - 1\right)^2 + \left(\frac{\sigma_X}{\sigma_Y} - 1\right)^2} \quad (69)$$

As may be seen in Equation 69, KGE penalizes deficient linear association, relative mean error and relative error in centred second-order moment. Unlike MSE, however, KGE is not strictly proper. Hence, one may game forecasts and forecast systems to score well (or poorly) on KGE. The above may be illustrated via simple examples. Under Equation 69, white-noise forecasts with $\rho_{X,Y} = 0$, $\mu_X = \mu_Y$ and $\sigma_X = \sigma_Y$, or forecast A, have a KGE of 0. Similarly, clairvoyant (but biased by a multiplicative factor) forecasts with $\rho_{X,Y} = 1$, $\mu_X = \left(1 + \sqrt{0.5}\right)\mu_Y$ and $\sigma_X = \left(1 + \sqrt{0.5}\right)\sigma_Y$, or forecast B, also have a KGE of 0. Hence, against intuition, the two forecasts are of equal "goodness" according to KGE. Using Equation 68, it is easy to show that $MSE_B > MSE_A$ if $CV_O < 1/\sqrt{3}$ (≈ 0.58) and $MSE_B < MSE_A$ otherwise, where MSE_A and MSE_B denote the MSE of forecast A and forecast B, respectively, and CV_Y denotes the coefficient of variation of the observation $(= \sigma_Y/\mu_Y)$. In other words, if the predictand has a small/large variability relative to the mean, biases are more/less important than association, in agreement with intuition. Equation 69, on the other hand, gives $KGE_A = KGE_B$ regardless of CV_Y. The above indicates that the same value of KGE does not, in general, mean the same level of forecast quality when comparing forecasts across different locations, seasons, flow regimes, and so on. Hence, if KGE is to be used for verification, it is recommended that it be used jointly with the MSE, its decomposition and the (R)MSE skill score.

Reliability, resolution and uncertainty

The CR decomposition expresses the MSE in terms of reliability (REL), resolution (RES) and uncertainty (UNC) as follows under the assumption of $\mu_X = \mu_Y$:

$$MSE = REL - RES + UNC = E[(X-Y)^2] = E_X[(\mu_{Y|X} - X)^2] - E_X[(\mu_{Y|X} - \mu_Y)^2] + \sigma_Y^2 \quad (70)$$

where $E_X[\ ]$ indicates that the expectation is with respect to the forecast, X, and $\mu_{Y|X}$ denotes the conditional expectation of the observation, Y, given the forecast, X. Note that $\mu_{Y|X}$ is a function of the random variable, X, and hence a random variable itself.

Figure 73 illustrates the terms in Equation 70. The circles connected by lines represent $\mu_{Y|X=x}$ for selected experimental values of X (that is, forecast events) on the x-axis. The horizontal line indicates μ_Y. The diagonal line represents $y = x$ (that is, the observed outcome is the same as the forecast event). The histogram above the large box represents the distribution of the forecast, $p_X(x)$. The reliability component REL, which is contributed by the red shaded area, measures the squared deviations from the diagonal line weight-averaged by $p_X(x)$ (see the first term in Equation 69). The forecast is perfectly reliable (REL = 0) if all $\mu_{Y|X=x}$ points fall on the diagonal line.

The resolution component RES, which is contributed by the blue shaded area, measures the squared deviations of $\mu_{Y|X=x}$ from the horizontal line representing μ_Y, weight-averaged by $p_X(x)$ (see the second term in Equation 70). In the example shown in Figure 73, the forecast has positive resolution since $\mu_{Y|X=x}$ shows sensitivity to x, indicating that the observed outcomes are different for different forecast events. If $\mu_{Y|X=x}$ is the same for all forecast events, the forecast has no resolution (RES = 0). The maximum resolution for perfectly reliable forecasts is equal to σ_Y^2, which would only be achieved if the forecast is perfectly accurate (MSE = 0). The MSE decreases as resolution increases. Hence, the forecasts must have good resolution to be accurate. The uncertainty component UNC is given by σ_Y^2, representing the variability of the observation about its mean. Climatological forecast is perfectly reliable (REL = 0) but has no resolution (RES = 0), and hence the MSE is equal to the variability of the observation, σ_Y^2.

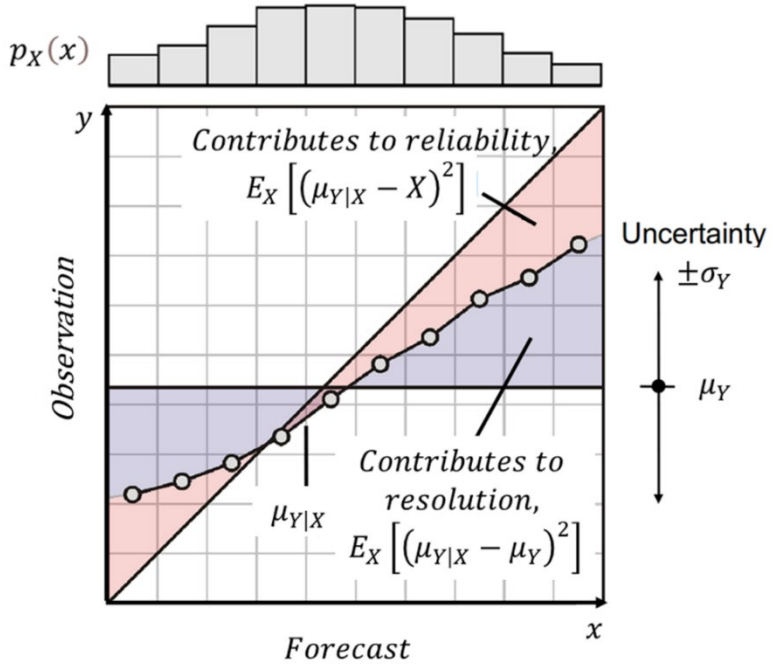


Figure 74. Conceptual illustration of the CR decomposition of the MSE of the forecast, x, versus the observation, y

Source: Adapted from Bradley et al. (2019)

Type II conditional bias, discrimination and sharpness

The LBR decomposition expresses the MSE in terms of type II conditional bias (T2B), discrimination (DIS) and sharpness (SHA) as follows under the assumption of $\mu_X = \mu_Y$:

$$MSE = T2B - DIS + SHA = E[(X - Y)^2] = E_Y[(\mu_{X|Y} - Y)^2] - E_Y[(\mu_{X|Y} - \mu_X)^2] + \sigma_X^2 \quad (71)$$

where E_Y indicates that the expectation is with respect to the observation, Y, and $\mu_{X|Y}$ denotes $E[X|Y]$. As with $\mu_{Y|X}$ in Equation 70, $\mu_{X|Y}$ in Equation 71 is a function of the random variable, Y, and hence a random variable itself.

Figure 74 is completely analogous to Figure 73 but illustrates the LBR decomposition. The large box is the same as in Figure 73. The circles connected by lines represent $\mu_{X|Y=y}$ for all different observed events. The vertical line indicates the mean of the forecast, μ_X. The histogram next to the box represents the marginal distribution of the observation, $p_Y(y)$. If $\mu_{X|Y=y}$ falls on the diagonal line for all observed events, the forecast has no type II conditional bias (T2B = 0). The type II conditional bias component T2B is given by the squared deviations from the diagonal line weight-averaged by the marginal distribution $p_Y(y)$ (see the first term in Equation 71) and is represented by the red shaded area.

The discrimination component, DIS, which is indicated by the blue shaded area, measures the squared deviations of $\mu_{X|Y=y}$ from the vertical line μ_X weight-averaged by the marginal distribution $p_Y(y)$ (see the second term in Equation 71). In the example shown in Figure 75, the forecast has positive discrimination; $\mu_{X|Y=y}$ increases as the observed event y increases, indicating that the forecasts are different for different observed events. The sharpness component SHA is represented by variance of the forecast, σ_X^2, which represents the variability of the forecast about its mean. An accurate forecast must have nonzero SHA to have positive DIS and small type II conditional bias. If the same forecast is always issued, the forecast has zero SHA, zero DIS and maximum type II conditional bias.

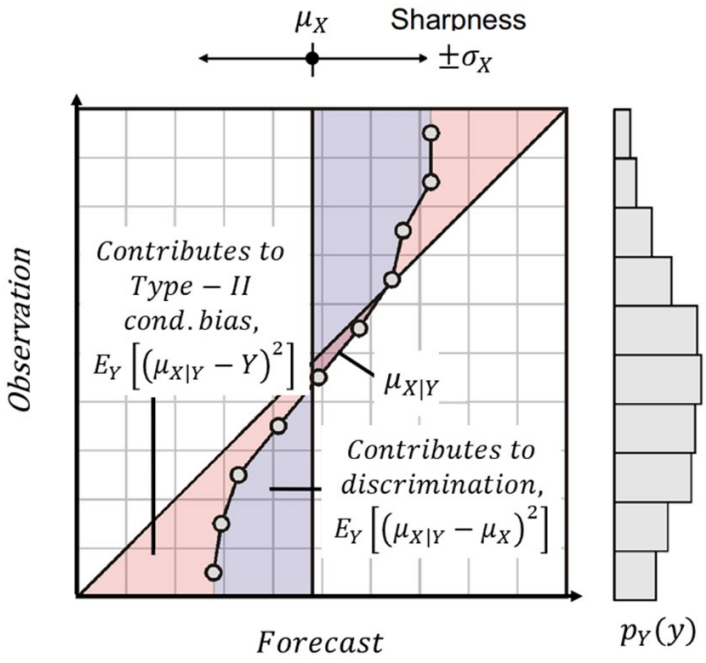


Figure 75. Conceptual illustration of the LBR decomposition

Source: Adapted from Bradley et al. (2019)

APPENDIX C. LIST OF ACRONYMS

BC	Boundary condition
BIAS	Multiplicative bias
BR	Base rate
BS	Brier score
BSS	Brier skill score
CDF	Cumulative distribution function
CORR	Pearson correlation coefficient
CR	Calibration–resolution
CRPS	Continuous ranked probability score
CRPSS	Continuous ranked probability skill score
CSI	Critical success index
DA	Data assimilation
DIS	Discrimination
DO	Distributions-oriented
DRB	Delaware River Basin
ECCC	Environment and Climate Change Canada
EnsPost	Ensemble Postprocessor
ERRIS	Error Reduction and Representation in Stages (error model)
ETS	Equitable threat score
EVS	Ensemble Verification System
FA	False alarm
FAR	False alarm ratio
FB	Frequency bias
FC	Fraction correct
GDPS	Global Deterministic Prediction System
GEFS	Global Ensemble Forecast System
GEPS	Global Ensemble Prediction System
H	Hit

HEFS	Hydrologic Ensemble Forecast Service
HyFS	Hydrological Forecasting System
IC	Initial condition
IID	Independent and identically distributed
KGE	Kling–Gupta efficiency
LBR	Likelihood–base rate
M	Miss
MAE	Mean absolute error
MAP	Mean areal precipitation
MARFC	Middle Atlantic River Forecast Center
ME	Mean error
MEFP	Meteorological Ensemble Forecast Processor
MSE	Mean squared error
MSESS	Mean squared error skill score
NAEFS	North American Ensemble Forecast System
NAM	North American Mesoscale Forecast System
NCEP	National Centers for Environmental Prediction
NRC	National Research Council
NWM	National Water Model
NWP	Numerical weather prediction
NWS	National Weather Service
OHRFC	Ohio River Forecast Center
PDF	Probability density function
PIT	Probability integral transform
PMF	Probability mass function
POD	Probability of detection
POFD	Probability of false detection
POFO	Probability of forecast of occurrence
PSS	Peirce’s skill score
QPF	Quantitative precipitation forecast

RDPS	Regional Deterministic Prediction System
REL	Reliability
RES	Resolution
RFC	River Forecast Center
RME	Relative mean error
RMSE	Root mean squared error
ROC	Relative operating characteristic
RPS	Ranked probability score
RPSS	Ranked probability skill score
RWsOS	Rijkswaterstaat Operational System
SAC	Sacramento soil moisture accounting model
SHA	Sharpness
SPH	Système de Prévision Hydrologique
SR	Success ratio
T2B	Type II bias
TN	True negative
UHG	Unit hydrograph
UNC	Uncertainty
UTRB	Upper Trinity River Basin
WMCN	Water Management Centre of the Netherlands
WWRP	World Weather Research Programme

REFERENCES

Alizadeh, B.; Limon, R. A.; Seo, D.-J. et al. Multiscale Postprocessor for Ensemble Streamflow Prediction for Short to Long Ranges. **2020**. https://doi.org/10.1175/JHM-D-19-0164.1.

Anctil, F.; Ramos, M.-H. Verification Metrics for Hydrological Ensemble Forecasts. In *Handbook of Hydrometeorological Ensemble Forecasting*; Duan, Q., Pappenberger, F., Wood, A., Cloke, H. L., Schaake, J. C., Eds.; Springer: Berlin, Heidelberg, 2019; 893–922. https://doi.org/10.1007/978-3-642-39925-1_3.

Anderson, S. R.; Csima, G.; Moore, R. J. et al. Towards Operational Joint River Flow and Precipitation Ensemble Verification: Considerations and Strategies given Limited Ensemble Records. *Journal of Hydrology* **2019**, *577*, 123966. https://doi.org/10.1016/j.jhydrol.2019.123966.

Azevedo, R.; Bernard, R. M. A Meta-Analysis of the Effects of Feedback in Computer-Based Instruction. *Journal of Educational Computing Research* **1995**, *13* (2), 111–127. https://doi.org/10.2190/9LMD-3U28-3A0G-FTQT.

Bellier, J.; Zin, I.; Bontron, G. Sample Stratification in Verification of Ensemble Forecasts of Continuous Scalar Variables: Potential Benefits and Pitfalls. **2017**, *145* (9), 3529–3544. https://doi.org/10.1175/MWR-D-16-0487.1.

Benedetti, R. Scoring Rules for Forecast Verification. *Monthly Weather Review* **2010**, *138* (1), 203–211. https://doi.org/10.1175/2009MWR2945.1.

Benjamin, J. R.; Cornell, C. A. *Probability, Statistics, and Decision for Civil Engineers*; Dover, 1970.

Bennett, J.; Robertson, D. Data for Paper "Propagating Reliable Estimates of Hydrological Forecast Uncertainty to Many Lead Times"; Commonwealth Scientific and Industrial Research Organisation (CSIRO), 2021. https://doi.org/10.25919/PB88-7824.

Bennett, J. C.; Robertson, D. E.; Wang, Q. J. et al. Propagating Reliable Estimates of Hydrological Forecast Uncertainty to Many Lead Times. *Journal of Hydrology* **2021**, *603*, 126798. https://doi.org/10.1016/j.jhydrol.2021.126798.

Bowler, N. E. Accounting for the Effect of Observation Errors on Verification of MOGREPS. *Meteorological Applications* **2008**, *15* (1), 199–205. https://doi.org/10.1002/met.64.

Bradley, A. A.; Demargne, J.; Franz, K. J. Attributes of Forecast Quality. In *Handbook of Hydrometeorological Ensemble Forecasting*; Duan, Q., Pappenberger, F., Wood, A., Cloke, H. L., Schaake, J. C., Eds.; Springer: Berlin, Heidelberg, 2019; 849–892. https://doi.org/10.1007/978-3-642-39925-1_2.

Bras, R. L.; Rodríguez-Iturbe, I. *Random Functions and Hydrology*; Addison-Wesley, 1984.

Breusch, T. S.; Pagan, A. R. A Simple Test for Heteroscedasticity and Random Coefficient Variation. *Econometrica* **1979**, *47* (5), 1287–1294. https://doi.org/10.2307/1911963.

Brier, G. W. Verification of Forecasts Expressed in Terms of Probability. *Monthly Weather Review* **1950**, *78* (1), 1–3. https://journals.ametsoc.org/view/journals/mwre/78/1/1520-0493_1950_078_0001_vofeit_2_0_co_2.xml.

Brier, G. W.; Allen, R. A. Verification of Weather Forecasts. In *Compendium of Meteorology: Prepared under the Direction of the Committee on the Compendium of Meteorology*; Byers, H. R.; Landsberg, H. E.; Wexler, H. et al., Eds.; American Meteorological Society: Boston, MA, 1951; 841–848. https://doi.org/10.1007/978-1-940033-70-9_68.

Brown, J. D.; Demargne, J.; Seo, D.-J. et al. The Ensemble Verification System (EVS): A Software Tool for Verifying Ensemble Forecasts of Hydrometeorological and Hydrologic Variables at Discrete Locations. *Environmental Modelling & Software* **2010**, *25* (7), 854–872. https://doi.org/10.1016/j.envsoft.2010.01.009.

Brown, J. D.; He, M.; Regonda, S. et al. Verification of Temperature, Precipitation, and Streamflow Forecasts from the NOAA/NWS Hydrologic Ensemble Forecast Service (HEFS): 2. Streamflow Verification. *Journal of Hydrology* **2014a**, *519*, 2847–2868. https://doi.org/10.1016/j.jhydrol.2014.05.030.

Brown, J. D.; Wu, L.; He, M. et al. Verification of Temperature, Precipitation, and Streamflow Forecasts from the NOAA/NWS Hydrologic Ensemble Forecast Service (HEFS): 1. Experimental Design and Forcing Verification. *Journal of Hydrology* **2014b**, *519*, 2869–2889. https://doi.org/10.1016/j.jhydrol.2014.05.028.

Brussolo, E.; Hardenberg, J. von; Ferraris, L. et al. Verification of Quantitative Precipitation Forecasts via Stochastic Downscaling. *Journal of Hydrometeorology* **2008**, *9* (5), 1084–1094. https://doi.org/10.1175/2008JHM994.1.

Burnash, R. J. C.; Ferral, R. L.; McGuire, R. A. *A Generalized Streamflow Simulation System – Conceptual Modeling for Digital Computers*; Joint Federal-State River Forecast Center: Sacramento, USA, 1973.

Burton, J. Robert FitzRoy and the Early History of the Meteorological Office. *The British Journal for the History of Science* **1986**, *19* (2), 147–176. https://doi.org/10.1017/S0007087400022949.

Casati, B.; Wilson, L. J.; Stephenson, D. B. et al. Forecast Verification: Current Status and Future Directions. *Meteorological Applications* **2008**, *15* (1), 3–18. https://doi.org/10.1002/met.52.

Chang, W.; Cheng, J.; Allaire. J. et al. *Shiny: Web Application Framework for R*; R package version 1.7.4.9001; Shiny, 2023.

Chow, V. T., Maidment, D. R.; Mays, L. W. Applied Hydrology; McGraw-Hill: New York, USA, 1988.

Cui, B.; Toth, Z.; Zhu, Y.; Hou, D. Bias Correction for Global Ensemble Forecast. *Weather and Forecasting* **2012**, *27* (2), 396-410. https://doi.org/10.1175/WAF-D-11-00011.1

Daan, H. *Scoring Rules in Forecast Verification* (Programme on Short- and Medium-range Weather Prediction Research Publication Series No. 4); World Meteorological Organization (WMO): Geneva, 1984.

Day, G. N. Extended Streamflow Forecasting Using NWSRFS. *Journal of Water Resources Planning and Management* **1985**, *111* (2), 157–170. https://doi.org/10.1061/(ASCE)0733-9496(1985)111:2(157).

Demargne, J.; Mullusky, M.; Werner, K. et al. Application of Forecast Verification Science to Operational River Forecasting in the U.S. National Weather Service. *Bulletin of the American Meteorological Society* **2009**, *90* (6), 779–784. https://doi.org/10.1175/2008BAMS2619.1.

Demargne, J.; Wu, L.; Regonda, S. K. et al. The Science of NOAA's Operational Hydrologic Ensemble Forecast Service. *Bulletin of the American Meteorological Society* **2014**, *95* (1), 79–98. https://doi.org/10.1175/BAMS-D-12-00081.1.

Di Baldassarre, G.; Montanari, A. Uncertainty in River Discharge Observations: A Quantitative Analysis. *Hydrology and Earth System Sciences* **2009**, *13* (6), 913–921. https://doi.org/10.5194/hess-13-913-2009.

Drake, A. W. *Fundamentals of Applied Probability Theory*; McGraw-Hill: New York, USA, 1967.

Duan, Q.; Schaake, J.; Andréassian, V. et al. Model Parameter Estimation Experiment (MOPEX): An Overview of Science Strategy and Major Results from the Second and Third Workshops. *Journal of Hydrology* **2006**, *320* (1), 3–17. https://doi.org/10.1016/j.jhydrol.2005.07.031.

Efron, B. Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics* **1979**, *7* (1), 1–26.

Finley, J. P. Tornado Predictions. *American Meteorological Journal* **1884**, *1*, 85–88.

Georgakakos, K. P.; Seo, D.-J.; Gupta, H. et al. Towards the Characterization of Streamflow Simulation Uncertainty through Multimodel Ensembles. *Journal of Hydrology* **2004**, *298* (1), 222–241. https://doi.org/10.1016/j.jhydrol.2004.03.037.

Gneiting, T.; Raftery, A. E. Strictly Proper Scoring Rules, Prediction, and Estimation. *Journal of the American Statistical Association* **2007**, *102* (477), 359–378. https://doi.org/10.1198/016214506000001437.

Gneiting, T.; Raftery, A. E.; Westveld, A. H. et al. Calibrated Probabilistic Forecasting Using Ensemble Model Output Statistics and Minimum CRPS Estimation. *Monthly Weather Review* **2005**, *133* (5), 1098–1118. https://doi.org/10.1175/MWR2904.1.

Gneiting, T.; Ranjan, R. Comparing Density Forecasts Using Threshold- and Quantile-Weighted Scoring Rules. *Journal of Business & Economic Statistics* **2011**, *29*(3), 411–422. https://doi.org/10.1198/jbes.2010.08110.

Gneiting, T.; Balabdaoui, F.; Raftery, A. E. Probabilistic Forecasts, Calibration and Sharpness. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **2007**, *69* (2), 243–268. https://doi.org/10.1111/j.1467-9868.2007.00587.x.

Graziano, T., E.; Clark, B.; Cosgrove et al. Transforming National Oceanic and Atmospheric Administration (NOAA) Water Resources Prediction. In *Abstracts of the 31st Conference on Hydrology*, Seattle, WA, USA, 23–26 January 2017; American Meteorological Society, 2A.2, 2017. https://ams.confex.com/ams/97Annual/webprogram/Paper314016.html.

Green, D. M.; Swets, J. A. Signal Detection Theory and Psychophysics; John Wiley and Sons: New York, USA, 1966.

Guan, H.; Zhu, Y.; Sinsky, E. et al. *The NCEP GEFS-v12 Reforecasts to Support Subseasonal and Hydrometeorological Applications*, Science and Technology Infusion Climate Bulletin – NOAA's National Weather Service, 44th NOAA Annual Climate Diagnostics and Prediction Workshop, Durham, USA, 22–24 October 2019.

Gupta, H. V.; Kling, H.; Yilmaz, K. K. et al. Decomposition of the Mean Squared Error and NSE Performance Criteria: Implications for Improving Hydrological Modelling. *Journal of Hydrology* **2009**, *377* (1), 80–91. https://doi.org/10.1016/j.jhydrol.2009.08.003.

Hamill, T. M. Interpretation of Rank Histograms for Verifying Ensemble Forecasts. *Monthly Weather Review* **2001**, *129* (3), 550–560. https://journals.ametsoc.org/view/journals/mwre/129/3/1520-0493_2001_129_0550_iorhfv_2.0.co_2.xml.

Hamill, T. M.; Juras, J. Measuring Forecast Skill: Is It Real Skill or Is It the Varying Climatology? *Quarterly Journal of the Royal Meteorological Society* **2006**, *132* (621C), 2905–2923. https://doi.org/10.1256/qj.06.25.

Harmel, R. D.; Cooper, R. J.; Slade, R. M. et al. Cumulative Uncertainty in Measured Streamflow and Water Quality Data for Small Watersheds. *Transactions of the ASABE* **2006**, *49* (3), 689–701.

Harris, D.; Foufoula-Georgiou, E.; Droegemeier, K. K. et al. Multiscale Statistical Properties of a High-Resolution Precipitation Forecast. *Journal of Hydrometeorology* **2001**, *2* (4),406–418. https://journals.ametsoc.org/view/journals/hydr/2/4/1525-7541_2001_002_0406_mspoah_2_0_co_2.xml.

Hersbach, H. Decomposition of the Continuous Ranked Probability Score for Ensemble Prediction Systems. *Weather and Forecasting* **2000**, *15* (5), 559–570. https://doi.org/10.1175/1520-0434(2000)015<0559:DOTCRP>2.0.CO;2.

Hsu, W.; Murphy, A. H. The Attributes Diagram A Geometrical Framework for Assessing the Quality of Probability Forecasts. *International Journal of Forecasting* **1986**, *2* (3), 285–293. https://doi.org/10.1016/0169-2070(86)90048-8.

Iliopoulos, G.; Karlis, D.; Ntzoufras, I. Bayesian Estimation in Kibble's Bivariate Gamma Distribution. *Canadian Journal of Statistics* **2005**, *33* (4), 571–589. https://doi.org/10.1002/cjs.5550330408.

Imhoff, R. O.; Brauer, C. C.; Overeem, A. et al. Spatial and Temporal Evaluation of Radar Rainfall Nowcasting Techniques on 1,533 Events. *Water Resources Research* **2020**, *56* (8), e2019WR026723. https://doi.org/10.1029/2019WR026723.

Jewson, S. The Problem with the Brier Score. *arXiv* **2004**. https://arxiv.org/abs/physics/0401046.

Johnson, R. A.; Evans, J. W.; Green, D. W. *Some Bivariate Distributions for Modeling the Strength Properties of Lumber*; Research Paper FPL–RP–575; USDA Forest Service Forest Products Laboratory: Madison, USA, 1999.

Jolliffe, I. T.; Stephenson, D. B. *Forecast Verification: A Practitioner's Guide in Atmospheric Science*; Wiley: 2012.

Jozaghi, A.; Shen, H.; Ghazvinian, M. et al. Multi-Model Streamflow Prediction Using Conditional Bias-Penalized Multiple Linear Regression. *Stochastic Environmental Research and Risk Assessment* **2021**, *35* (11), 2355–2373. https://doi.org/10.1007/s00477-021-02048-3.

Keller, J. D.; Hense, A. A New Non-Gaussian Evaluation Method for Ensemble Forecasts Based on Analysis Rank Histograms. *Meteorologische Zeitschrift* **2011**, *20* (2), 107–117. https://doi.org/10.1127/0941-2948/2011/0217.

Kim, S.; Seo, D.-J.; Riazi, H. et al. Improving Water Quality Forecasting via Data Assimilation – Application of Maximum Likelihood Ensemble Filter to HSPF. *Journal of Hydrology* **2014**, *519*, 2797–2809. https://doi.org/10.1016/j.jhydrol.2014.09.051.

Kim, S.; Sadeghi, H.; Limon, R. A. et al. Assessing the Skill of Medium-Range Ensemble Precipitation and Streamflow Forecasts from the Hydrologic Ensemble Forecast Service (HEFS) for the Upper Trinity River Basin in North Texas. *Journal of Hydrometeorology* **2018**, *19* (9), 1467–1483. https://doi.org/10.1175/JHM-D-18-0027.1.

Kim, S.; Shen, H.; Noh, S. et al. High-Resolution Modeling and Prediction of Urban Floods Using WRF-Hydro and Data Assimilation. *Journal of Hydrology* **2021**, *598*, 126236. https://doi.org/10.1016/j.jhydrol.2021.126236.

Krzysztofowicz, R. Bayesian Theory of Probabilistic Forecasting via Deterministic Hydrologic Model. *Water Resources Research* **1999**, *35* (9), 2739–2750. https://doi.org/10.1029/1999WR900099.

Kulik, J. A.; Kulik, C.-L. C. Timing of Feedback and Verbal Learning. *Review of Educational Research* **1988**, *58* (1), 79–97. https://doi.org/10.3102/00346543058001079.

Laio, F.; Tamea, S. Verification Tools for Probabilistic Forecasts of Continuous Hydrological Variables. *Hydrology and Earth System Sciences* **2007**, *11* (4), 1267–1277. https://doi.org/10.5194/hess-11-1267-2007.

Laugesen, R.; Thyer, M.; McInerney, D. et al. Flexible Forecast Value Metric Suitable for a Wide Range of Decisions: Application Using Probabilistic Subseasonal Streamflow Forecasts. *Hydrology and Earth System Sciences* **2023**, *27* (4), 873–893. https://doi.org/10.5194/hess-27-873-2023.

Lerch, S.; Thorarinsdottir, T. L.; Ravazzolo, F. et al. Forecaster's Dilemma: Extreme Events and Forecast Evaluation. *Statistical Science* **2017**, *32* (1), 106–127. https://doi.org/10.1214/16-STS588.

Li, M.; Wang, Q. J.; Bennett, J. C. et al. A Strategy to Overcome Adverse Effects of Autoregressive Updating of Streamflow Forecasts. *Hydrology and Earth System Sciences* **2015**, *19* (1), 1–15. https://doi.org/10.5194/hess-19-1-2015.

Lieberman, M. D.; Cunningham, W. A. Type I and Type II Error Concerns in fMRI Research: Re-Balancing the Scale. *Social Cognitive and Affective Neuroscience* **2009**, *4* (4), 423–428. https://doi.org/10.1093/scan/nsp052.

Linsley, R. K.; Kohler, M. A.; Paulhus, J. L. H. *Hydrology for Engineers*; 3rd ed.; New York, USA, McGraw-Hill: 1982.

Liu, Y.; Brown, J.; Demargne, J. et al. A Wavelet-Based Approach to Assessing Timing Errors in Hydrologic Predictions. *Journal of Hydrology* **2011**, *397* (3), 210–224. https://doi.org/10.1016/j.jhydrol.2010.11.040.

Liu, Y.; Weerts, A. H.; Clark, M. et al. Advancing Data Assimilation in Operational Hydrologic Forecasting: Progresses, Challenges, and Emerging Opportunities. *Hydrology and Earth System Sciences* **2012**, *16* (10), 3863–3887. https://doi.org/10.5194/hess-16-3863-2012.

Lu, J.-C.; Bhattacharyya, G. K. Some New Constructions of Bivariate Weibull Models. *Ann Inst Stat Math* **1990**, *42* (3), 543–559. https://doi.org/10.1007/BF00049307.

Matheson, J. E.; Winkler, R. L. Scoring Rules for Continuous Probability Distributions. *Management Science* **1976**, *22* (10), 1087–1096.

Milly, P. C. D.; Betancourt, J.; Falkenmark, M. et al. Stationarity Is Dead: Whither Water Management? *Science* **2008**, *319* (5863), 573–574. https://doi.org/10.1126/science.1151915.

Moriasi, D. N.; Gitau, M. W.; Pai, N. et al. Hydrologic and Water Quality Models: Performance Measures and Evaluation Criteria. *Transactions of the ASABE* **2015**, *58*(6), 1763–1785. https://doi.org/10.13031/trans.58.10715.

Murphy, A. H. On Expected-Utility Measures in Cost-Loss Ratio Decision Situations. *Journal of Applied Meteorology and Climate* **1969**, *8* (6), 989–991. https://journals.ametsoc.org/view/journals/apme/8/6/1520-0450_1969_008_0989_oeumic_2_0_co_2.xml.

Murphy, A. H. A New Vector Partition of the Probability Score. *Journal of Applied Meteorology and Climatology* **1973**, *12* (4), 595–600. https://doi.org/10.1175/1520-0450(1973)012<0595:ANVPOT>2.0.CO;2.

Murphy, A. H. What Is a Good Forecast? An Essay on the Nature of Goodness in Weather Forecasting. *Weather and Forecasting* **1993**, *8* (2), 281–293. https://journals.ametsoc.org/view/journals/wefo/8/2/1520-0434_1993_008_0281_wiagfa_2_0_co_2.xml.

Murphy, A. H. The Finley Affair: A Signal Event in the History of Forecast Verification. *Weather and Forecasting* **1996**, *11* (1), 3–20. https://journals.ametsoc.org/view/journals/wefo/11/1/1520-0434_1996_011_0003_tfaase_2_0_co_2.xml.

Murphy, A. H. Forecast Verification. In *Economic Value of Weather and Climate Forecasts*, Katz, R. W.; Murphy, A. H., Eds.; Cambridge University Press, 1997; 19–74.

Murphy, A. H.; Winkler, R. L. A General Framework for Forecast Verification. *Monthly Weather Review* **1987**, *115* (7), 1330–1338. https://journals.ametsoc.org/view/journals/mwre/115/7/1520-0493_1987_115_1330_agfffv_2_0_co_2.xml.

Murphy, A. H.; Ehrendorfer, M. On the Relationship between the Accuracy and Value of Forecasts in the Cost–Loss Ratio Situation. *Weather and Forecasting* **1987**, *2* (3), 243–251. https://doi.org/10.1175/1520-0434(1987)002<0243:OTRBTA>2.0.CO;2.

Nagao, M.; Kadoya, M. The Study on Bivariate Gamma Distribution and its Applicability. *Annuals of the Disaster Prevention Research Institute of Kyoto University* **1970**, *13B*, 105–115.

Nash, J. E.; Sutcliffe, J. V. River Flow Forecasting through Conceptual Models Part I — A Discussion of Principles. *Journal of Hydrology* **1970**, *10* (3), 282–290. https://doi.org/10.1016/0022-1694(70)90255-6.

National Research Council (NRC). *Grand Challenges in Environmental Sciences*; National Academies Press: Washington, DC, USA, 2001. https://doi.org/10.17226/9975.

National Weather Service (NWS). *Inundation Mapping Release 2.0*; Office of Hydrologic Development, Office of Climate, Water and Weather Services, Silver Spring, USA, 2012.

National Weather Service (NWS). *Meteorological Ensemble Forecast Processor (MEFP) User's Manual*; Office of Hydrologic Development, NWS, 2017a. https://vlab.ncep.noaa.gov/documents/207461/1893026/MEFP_Users_Manual.pdf.

National Weather Service (NWS). *Ensemble Postprocessor (EnsPost) User's Manual*; Office of Hydrologic Development, NWS, 2017b. https://vlab.ncep.noaa.gov/documents/207461/1893026/EnsPost_Users_Manual.pdf.

National Center for Atmospheric Research (NCAR) – Research Applications Laboratory. Verification: Weather Forecast Verification Utilities; 2015. https://cran.r-project.org/web/packages/verification/index.html.

Nelson, B. R.; Seo, D.-J.; Kim, D. Multisensor Precipitation Reanalysis. *Journal of Hydrometeorology* **2010**, *11* (3), 666–682. https://doi.org/10.1175/2010JHM1210.1.

Noh, S. J.; Lee, J.-H.; Lee, S. et al. Retrospective Dynamic Inundation Mapping of Hurricane Harvey Flooding in the Houston Metropolitan Area Using High-Resolution Modeling and High-Performance Computing. *Water* **2019**, *11* (3), 597. https://doi.org/10.3390/w11030597.

Nipen, T. *WFRT Verif*; Github, 2016. https://github.com/WFRT/verif.

Oreskes, N.; Shrader-Frechette, K.; Belitz, K. Verification, Validation, and Confirmation of Numerical Models in the Earth Sciences. *Science* **1994**, *263* (5147), 641–646. https://doi.org/10.1126/science.263.5147.641.

Pagano, T.; Garen, D. A Recent Increase in Western U.S. Streamflow Variability and Persistence. *Journal of Hydrometeorology* **2005**, *6* (2), 173–179. https://doi.org/10.1175/JHM410.1.

Pappenberger, F.; Buizza, R. The Skill of ECMWF Precipitation and Temperature Predictions in the Danube Basin as Forcings of Hydrological Models. *Weather and Forecasting* **2009**, *24* (3), 749–766. https://doi.org/10.1175/2008WAF2222120.1.

Pappenberger, F.; Scipal, K.; Buizza, R. Hydrological Aspects of Meteorological Verification. *Atmospheric Science Letters* **2008**, *9* (2), 43–52. https://doi.org/10.1002/asl.171.

Pappenberger, F.; Ramos, M. H.; Cloke, H. L. et al. How Do I Know If My Forecasts Are Better? Using Benchmarks in Hydrological Ensemble Prediction. *Journal of Hydrology* **2015**, *522*, 697–713. https://doi.org/10.1016/j.jhydrol.2015.01.024.

Perica, S.; Pavlovic, S.; St. Laurent, M. et al. *NOAA Atlas 14: Precipitation-Frequency Atlas of the United States – Texas*; Volume 11; Version 2.0; National Weather Service, Silver Spring, USA, 2018.

Potts, J. M. Basic Concepts. In *Forecast Verification*; Jolliffe, I. T.; Stephenson, D. B., Eds.; John Wiley & Sons, 2011; 11–29.

Regonda, S. K.; Seo, D.-J.; Lawrence, B. et al. Short-Term Ensemble Streamflow Forecasting Using Operationally-Produced Single-Valued Streamflow Forecasts – A Hydrologic Model Output Statistics (HMOS) Approach. *Journal of Hydrology* **2013**, *497*, 80–96. https://doi.org/10.1016/j.jhydrol.2013.05.028.

Renard, B.; Kavetski, D.; Kuczera, G. et al. Understanding Predictive Uncertainty in Hydrologic Modeling: The Challenge of Identifying Input and Structural Errors. *Water Resources Research* **2010**, *46* (5). https://doi.org/10.1029/2009WR008328.

Reynolds, D. Value-added Quantitative Precipitation Forecasts: How Valuable Is the Forecaster? *Bulletin of the American Meteorological Society* **2003**, *84* (7), 876–878.

Riggs, H. C. Regional Analyses of Streamflow Characteristics, United States Government Printing Office: Washington, DC, USA, 1973.

Ritchie, H.; Samborska, V.; Roser, M. *Urbanization*. Our World in Data, 2018, updated in 2024. https://ourworldindata.org/urbanization.

Roebber, P. J. Visualizing Multiple Measures of Forecast Quality. *Weather and Forecasting* **2009**, *24* (2), 601-608. https://doi.org/10.1175/2008WAF2222159.1.

Schaake, J.; Demargne, J.; Hartman, R. et al. Precipitation and Temperature Ensemble Forecasts from Single-Value Forecasts. *Hydrology and Earth System Sciences Discussions* **2007**, *4* (2), 655–717. https://doi.org/10.5194/hessd-4-655-2007.

Searcy, J. K. *Flow-duration Curves*; No. 1542. United States Government Printing Office: Washington, DC, USA, 1959.

Seo, D.-J.; Herr, H. D.; Schaake, J. C. A Statistical Post-Processor for Accounting of Hydrologic Uncertainty in Short-Range Ensemble Streamflow Prediction. *Hydrology and Earth System Sciences Discussions* **2006**, *3* (4), 1987–2035. https://doi.org/10.5194/hessd-3-1987-2006.

Seo, D.-J.; Shen, H.; Lee, H. Adaptive Conditional Bias-Penalized Kalman Filter with Minimization of Degrees of Freedom for Noise for Superior State Estimation and Prediction of Extremes. *Computers & Geosciences* **2022**, *166*, 105193. https://doi.org/10.1016/j.cageo.2022.105193.

Shen, H.; Lee, H.; Seo, D.-J. Adaptive Conditional Bias-Penalized Kalman Filter for Improved Estimation of Extremes and Its Approximation for Reduced Computation. *Hydrology* **2022a**, *9* (2), 35. https://doi.org/10.3390/hydrology9020035.

Shen, H.; Seo, D.-J.; Lee, H. et al. Improving Flood Forecasting Using Conditional Bias-Aware Assimilation of Streamflow Observations and Dynamic Assessment of Flow-Dependent Information Content. *Journal of Hydrology* **2022b**, *605*, 127247. https://doi.org/10.1016/j.jhydrol.2021.127247.

Siddique, R.; Mejia, A.; Brown, J. et al. Verification of Precipitation Forecasts from Two Numerical Weather Prediction Models in the Middle Atlantic Region of the USA: A Precursory Analysis to Hydrologic Forecasting. *Journal of Hydrology* **2015**, *529*, 1390–1406. https://doi.org/10.1016/j.jhydrol.2015.08.042.

Stephenson, D. B.; Jolliffe, I. T. Forecast Verification: Past, Present and Future. In *Forecast Verification: A Practitioner's Guide in Atmospheric Science*; Jolliffe, I. T.; Stephenson, D. B., Eds.; John Wiley & Sons, 2003.

Talagrand, O. Assimilation of Observations, an Introduction (gtSpecial IssueltData Assimilation in Meteology and Oceanography: Theory and Practice). *Journal of the Meteorological Society of Japan* Ser. II **1997**, *75* (1B), 191–209. https://doi.org/10.2151/jmsj1965.75.1B_191.

Toth, Z.; Kalnay, E. Ensemble Forecasting at NCEP and the Breeding Method. *Monthly Weather Review* **1997**, *125* (12), 3297–3319. https://journals.ametsoc.org/view/journals/mwre/125/12/1520-0493_1997_125_3297_efanat_2.0.co_2.xml.

Troin, M.; Arsenault, R.; Wood, A. W. et al. Generating Ensemble Streamflow Forecasts: A Review of Methods and Approaches Over the Past 40 Years. *Water Resources Research* **2021**, *57* (7), e2020WR028392. https://doi.org/10.1029/2020WR028392.

Tufte, E. R. *The Visual Display of Quantitative Information*; 2nd ed.; 8th printing; Graphics Press: Cheshire, USA, 2013.

Welles, E.; Sorooshian, S.; Carter, G. et al. Hydrologic Verification: A Call for Action and Collaboration. *Bulletin of the American Meteorological Society* **2007**, *88* (4), 503–512. https://doi.org/10.1175/BAMS-88-4-503.

Wells, E. *National Weather Service (NWS): Hydrologic Ensemble Forecast Service*; NWS: 2017. https://www.weather.gov/media/wrn/calendar/HydrologicEnsembleForecastingService.pdf.

Werner, K.; Verkade, J. S.; Pagano, T. C. Application of Hydrological Forecast Verification Information. In *Handbook of Hydrometeorological Ensemble Forecasting*; Duan, Q., Pappenberger, F., Wood, A. et al., Eds.; Springer: Berlin, Heidelberg, 2019; 1013–1033. https://doi.org/10.1007/978-3-642-39925-1_7.

White, K. Delay of Test Information Feedback and Learning in a Conventional Classroom. *Psychology in the Schools* **1968**, *5* (1), 78–81. https://doi.org/10.1002/1520-6807(196801)5:1<78::AID-PITS2310050113>3.0.CO;2-Q.

Wilks, D. S. *Statistical Methods in the Atmospheric Sciences*, 3rd ed.; Academic Press, 2011.

Wilson, L. *Is there a Difference between "Verification" and "Validation"? Some Frequently Asked Questions*; WWRP/WGNE Joint Working Group on Forecast Verification Research, 2017. https://www.cawcr.gov.au/projects/verification/.

World Meteorological Organization (WMO). *Valuing Weather and Climate: Economic Assessment of Meteorological and Hydrological Services* (WMO-No. 1153); Geneva, 2015.

Wu, L.; Seo, D.-J.; Demargne, J. et al. Generation of Ensemble Precipitation Forecast from Single-Valued Quantitative Precipitation Forecast for Hydrologic Ensemble Prediction. *Journal of Hydrology* **2011**, *399* (3), 281–298. https://doi.org/10.1016/j.jhydrol.2011.01.013.

Zappa, M.; Fundel, F.; Jaun, S. A 'Peak-Box' Approach for Supporting Interpretation and Verification of Operational Ensemble Peak-Flow Forecasts. *Hydrological Processes* **2013**, *27* (1), 117–131. https://doi.org/10.1002/hyp.9521.

Zhou, X.; Zhu, Y.; Hou, D. et al. Performance of the New NCEP Global Ensemble Forecast System in a Parallel Experiment. *Weather and Forecasting* **2017**, *32* (5), 1989–2004. https://doi.org/10.1175/WAF-D-17-0023.1.

Zhu, Y.; Toth, Z. *NAEFS and NCEP Global Ensemble*; Presentation for National DOH Workshop; United States National Weather Service, 2008. https://www.nws.noaa.gov/oh/hrl/hsmb/docs/hep/events_announce/NAEFS_Yuejian_Zhu.pdf.

Zhu, Y.; Toth, Z.; Wobus, R. et al. The Economic Value of Ensemble-based Weather Forecasts. *Bulletin of the American Meteorological Society* **2002**, *83* (1), 73–84. https://journals.ametsoc.org/view/journals/bams/83/1/1520-0477_2002_083_0073_tevoeb_2_3_co_2.xml.

For more information, please contact:

World Meteorological Organization

7 bis, avenue de la Paix – P.O. Box 2300 – CH 1211 Geneva 2 – Switzerland

Strategic Communications Office

Tel.: +41 (0) 22 730 83 14

Email: cpa@wmo.int

wmo.int

JN 241123