Guidelines on Ensemble Prediction System Postprocessing

2021 edition

WEATHER CLIMATE WATER

WORLD
METEOROLOGICAL
ORGANIZATION

WMO-No. 1254

Guidelines on Ensemble Prediction System Postprocessing

2021 edition

WORLD
METEOROLOGICAL
ORGANIZATION

WMO-No. 1254

WMO-No. 1254

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
Fax: +41 (0) 22 730 81 17
Email: publications@wmo.int

ISBN 978-92-63-11254-5

PUBLICATION REVISION TRACK RECORD

Date	*Part/chapter/ section*	*Purpose of amendment*	*Proposed by*	*Approved by*

CONTENTS

ACKNOWLEDGEMENTS

This publication was prepared by the WMO Commission for Observation, Infrastructure and Information Systems (INFCOM) Standing Committee on Data Processing for Applied Earth System Modelling and Prediction (SC-ESMP) through concerted efforts by the following authors:

Ken MYLNE, Met Office, United Kingdom of Great Britain and Northern Ireland
Jing CHEN, China Meteorological Administration, China
Amin ERFANI, Meteorological Service of Canada, Canada
Tom HAMILL, National Oceanic and Atmospheric Administration (NOAA), United States of America
David RICHARDSON, European Centre for Medium-Range Weather Forecasts (ECMWF)
Stéphane VANNITSEM, Royal Meteorological Institute of Belgium, Belgium
Yong WANG, Central Institute of Meteorology and Geodynamics (ZAMG), Austria
Yuejian ZHU, National Oceanic and Atmospheric Administration, USA

This publication was reviewed by:

Hamza Athumani KABELWA, Tanzania Meteorological Agency, United Republic of Tanzania
Stephanie LANDMAN, South African Weather Service, South Africa
Eunha LIM, WMO
Jianjie WANG, China Meteorological Administration, China

The following individuals also provided valuable contributions to this publication:

Aitor ATENCIA, ZAMG, Austria
Markus DABERNIG, ZAMG, Austria

EXECUTIVE SUMMARY

The present guidelines on ensemble prediction system (EPS) postprocessing describe various postprocessing methods by which WMO Members can use information from available EPS forecasts to enhance and improve forecasts for their own specific regions or areas. They provide background on which statistical methods and data choices may be used for training, real-time forecasting, and validation. This publication is not a comprehensive instruction manual on how to implement methods or an explanation of the detailed mathematics behind the methods used; however, wherever possible, it provides references to where such information can be found. References to where available postprocessing software can be found are also provided.

These guidelines cover aspects of both physical and statistical postprocessing and take into consideration the opportunities offered by data science methods. With respect to physical postprocessing, a number of aspects are examined, including meteorological diagnosis and orographic downscaling. For statistical postprocessing, issues covered include bias correction, deterministic model output statistics, and ensemble calibration. The use of verification techniques to test and validate the postprocessing of both deterministic and probabilistic (EPS) forecasts is also discussed.

The present guidelines propose that WMO Members access real-time forecast data, historical data and reforecast data sets from the WMO Global Data-processing and Forecasting System (GDPFS). Obtaining data from GDPFS is much more cost-effective for Members than independently operating their own numerical weather prediction (NWP) systems. Postprocessing can greatly enhance the accuracy of real-time forecast data for applications at relatively low cost. The development of many of the postprocessing methods requires access to historical and reforecast data, both for statistical training and for validation purposes.

KEY RECOMMENDATIONS

1. Statistical postprocessing has consistently been demonstrated to improve the quality of both ensemble and deterministic forecasts and is one of the most cost-effective ways to produce higher-quality products. It is recommended that National Meteorological and Hydrological Services (NMHSs) utilize these postprocessing methods to enhance their forecasting capabilities.

2. NMHSs can apply postprocessing methods to model data which are available from existing prediction centres at a minimal cost relative to the cost required to operate an NWP system. It is therefore strongly recommended that NMHSs leverage data from WMO-designated GDPFS centres (see Section 8.1).

3. An archive of quality-controlled observations and past forecasts is essential for the training of statistical postprocessing and data science techniques and for validation and verification purposes. It is recommended that NMHSs continue to archive local data and, where possible, that they share these data with Regional Specialized Meteorological Centres (RSMCs) and global centres for statistical adaptation and model calibration purposes.

4. When beginning to apply postprocessing methods, it is recommended that NMHSs start with simple variables, such as surface temperature, and with data from their own local stations, rather than gridded data (which require greater storage and computation capabilities).

5. For deterministic forecasts of easier variables, it is recommended that NMHSs start by using the decaying average bias correction method, also referred to as the Kalman filter-type method (see Section 3.1). For ensemble forecasts, it is recommended that NMHSs start by using the ensemble model output statistics (EMOS) method (see Section 3.3.1).

6. It is essential that NMHSs use best practices in the development of their postprocessing methods. These practices include separating training and validation data and verifying both the original and the postprocessed forecasts against the validation data using a number of metrics (for example, continuous ranked probability skill score (CRPSS), reliability) and by comparing a visualization of the forecast with the verifying observations. (Ideally, this should be done by independent forecasters). Languages such as R and Python, for which there are many available software packages, can aid in the development of postprocessing methods (see Section 8.3 and Section 8.4).

7. In order to construct an operational forecasting system, NMHSs must have the ability to periodically obtain model and observation/analysis data for training and validation. In addition, NMHSs must have the capability to develop appropriate postprocessing methods and must be able to display and validate the products of those methods. Implementing and maintaining a comprehensive operational forecasting system will entail a significant investment of time, resources and effort.

CHAPTER 1. INTRODUCTION

Numerical weather prediction (NWP) involves the use of mathematical models capable of simulating the atmosphere to forecast the development of the weather over the next several days. Models are initialized using an analysis of the current state of the atmosphere generated from the most recent observations through a process called data assimilation. There are two standard forms of NWP: the older deterministic method, and the newer ensemble prediction system (EPS) method. The deterministic method is still widely employed today and generates a single model forecast. The problem with deterministic forecasts, however, is that they provide no information on the confidence or uncertainty in a forecast. Sometimes, for example, a forecast can go badly wrong very quickly and the forecaster or user has no warning of this. An EPS, in contrast, runs the model many times instead of just once. Each forecast within an ensemble is referred to as an ensemble member, and the members are initiated from very slightly different versions of the analysis. If all the members evolve similarly, this provides high confidence in the forecast. If, however, the various members diverge from each other, this indicates less confidence in the forecast, with the ensemble providing an estimate of the probabilities of different outcomes. EPSs are therefore generally used to provide probabilistic forecasts and to support risk-based forecasts and warnings. EPSs have become a powerful tool for forecasting weather and its uncertainty across the globe and underpin severe weather warnings intended to protect life and property.

Developing, operating and regularly improving modern EPSs capable of producing high-quality data is extremely expensive. By comparison, postprocessing systems which enhance the quality of EPS forecasts are much more economical to develop and apply and can result in improvements in the skill of EPS forecasts equivalent to many years of EPS system upgrades, particularly for local weather forecasts.

A number of Global Data-processing and Forecasting System (GDPFS) centres designated by WMO as World Meteorological Centres (WMCs) or Regional Specialized Meteorological Centres (RSMCs) make NWP data (from both EPSs and individual deterministic models) available for use and postprocessing by WMO Members. Even for National Meteorological and Hydrological Services (NMHSs) with their own EPS or deterministic NWP systems, investing in postprocessing systems will greatly enhance the quality and usefulness of their forecasts.

These guidelines provide an overview of those postprocessing methods which have been proven to be effective. Although the focus is on EPSs, some simple deterministic postprocessing is also included as a starting point and to aid in understanding.

This publication does not provide a full documentation of postprocessing methods or a step-by-step guide to their implementation; however, it does provide references for further details. Two particular books are especially recommended in this regard: *Statistical Postprocessing of Ensemble Forecasts*, by S. Vannitsem, D.S. Wilks and J.K. Messner and *Statistical Methods in the Atmospheric Sciences* by D.S. Wilks (see the References section for full bibliographic details).

Postprocessing methods range from the relatively simple to the highly complex, and the computing resources, data and technical expertise of the staff required to develop and implement these methods are variable. In order to help WMO Members select the methods most suitable for their needs, capabilities and resources, each method described is allocated to one of the three tiers outlined in Table 1. In general, it is recommended that NMHSs should start by implementing a method corresponding to the complexity, requirements and limitations indicated for Tier 1. These methods are some of the simplest postprocessing methods and can produce substantial benefits while requiring relatively little investment. NMHSs can then progress to the more advanced methods described in Tiers 2 and 3 as requirements and resources allow.

Table 1. Breakdown of postprocessing methods according to complexity, requirements and limitations

Tier	*Scientific and mathematical complexity*	*Technical implementation complexity and resource demands*	*Data requirements*	*Limitations*
Tier 1	Conceptually intuitive systems and basic university-level mathematics	May be implemented at a basic level on a desktop PC; standard software likely to be available and implemented without specialist software engineering skills	Access to basic EPS (and individual deterministic model) outputs and observations or analyses with a short historical record (≤1 year); typically, single-site, rather than gridded data sets	Univariate improvements may lose consistency or covariances; may degrade forecasts of rare or extreme events
Tier 2	Moderate complexity, possibly involving multivariate inputs and some performance optimization requiring scientific testing skills before implementation	Requires a high-powered desktop PC or larger computer, high-bandwidth data connectivity and the ability to process large (Gbyte) data sets	Access to gridded EPS fields and/or long archive records of site data amounting to many Mbytes (up to one Gbyte); complex data handling to match observations to forecasts (for example, large databases)	Risk of over-fitting to rare and extreme events
Tier 3	Advanced mathematical methods and/or detailed meteorological understanding required	Requires high-performance computing installation and expert software engineering skills to implement operationally or for real-time prediction; may require access to large "big data" archives (many Gbytes) with high connectivity	May require very large data sets or long time series of data for training	Challenging to write algorithms and to put into operation

This table is intended to provide a general overview of the various tiers of postprocessing methods. Not all methods will clearly fit into a particular tier. For example, a certain method might meet the requirements for Tier 1 but might require Tier 2-level computational capabilities if applied to large numbers of sites or high-resolution gridded data. In general, a method will be categorized according to its highest demand with respect to its scientific, technical and data requirements.

Table 2, below, presents a specific example of the breakdown of the Kalman filter model output statistics (MOS) postprocessing method. The Kalman filter MOS postprocessing method is a Tier 1 method because its complexities, data requirements and limitations are consistent with those described in the Tier 1 category in Table 1. Throughout these guidelines, when a postprocessing method is introduced, the tier to which it corresponds will be indicated in bold.

Table 2. Example of the complexities, requirements and limitations of the Kalman filter MOS method

Tier	*Scientific and mathematical complexity*	*Technical implementation complexity and resource demands*	*Data requirements*	*Limitations*
Tier 1	Kalman filter MOS method for bias correction of site forecasts using site observations	Standard Python or R code from software library implemented on a PC	Site forecasts provided from an EPS centre for a small number of sites on a daily basis and received over a low bandwidth connection; site observations from a local archive updated daily	May degrade extreme forecasts

Chapter 2 in this publication describes simple postprocessing techniques which can add significant benefits in terms of increasing the accuracy of a forecast. These techniques include calculating additional diagnostics from a small number of model output variables based on a physical understanding of the geography and atmospheric conditions, selecting the most relevant values for a location from model grid values according to factors such as coastlines and land elevation, and adjusting the model values to account for elevation using simple physical laws. These techniques can be applied equally to deterministic model outputs or to each member of an EPS.

Where a set of past forecasts and observations is available for one or more locations and where errors are Gaussian-distributed and relatively consistent across samples, simple bias corrections or regression relations can be derived to correct systematic errors. However, more involved techniques are typically needed for elements such as heavy precipitation amount or precipitation type. Statistical methods which are designed to improve the quality of forecasts of a specific element of the weather are called "univariate", and candidate univariate techniques for both deterministic forecasts and full ensemble probability distributions are briefly described in Chapter 3.

Both EPSs and individual deterministic models provide gridded scenarios where many elements of the weather are physically consistent with other elements, both temporally and spatially. However, many univariate calibration methods do not take this consistency into account. For some applications, these covariances are important, and multivariate techniques which do account for physical consistencies among elements are described in Chapter 4.

Chapter 5 provides some brief guidance on suitable approaches to use with multi-model ensembles if one synthesized forecast is desired.

It is essential that forecasts be verified and validated in order to allow end users to obtain quantitative and qualitative information regarding forecast skill and to ensure that postprocessing methods are improving forecast accuracy as expected. Chapter 6 provides some background on common verification metrics and the best practices to follow when constructing a verification system. Key reference material includes the Australian website https://www.cawcr.gov.au/projects/verification/, the contents of which were developed by the WMO Joint Working Group on Forecast Verification Research and books by Wilks (2019) and Jolliffe and Stephenson (2012).

Chapter 7 discusses general issues concerning data handling and computational requirements. Statistical postprocessing methods depend on the availability of high-quality data, and this chapter reviews some of the issues which can be expected when obtaining and maintaining training and validation data. While it has been a common experience that short-lead forecasts of variables such as surface temperature can be improved using a relatively short time series of past forecasts and observations, both longer-lead forecasts and forecasts of more extreme events, such as heavy precipitation, may benefit from longer training data sets and data from more than one prediction system. Long-lead forecasts and extreme event forecasts are both in demand, the former to provide more lead time to make decisions and the latter because of the much greater societal impact of extreme events.

Chapter 8 contains technical information, such as where to obtain the software and data necessary to build or improve postprocessing systems.

CHAPTER 2. PHYSICAL POSTPROCESSING

An NWP model uses the principles of dynamics and physics in relation to the atmosphere to make predictions about weather. Depending on the model's horizontal grid spacing and number of vertical levels, it may not fully resolve all atmospheric processes at the grid scale. If this is the case, the NWP model needs to use physical parameterizations, such as condensation and convective schemes, at the sub-grid scale to fill in the gap. However, in some situations, even with the most sophisticated parameterization schemes, some of the finer details of the atmosphere are not predicted well. Examples include the mode and severity of convective storms and the partitioning between the various phases of the precipitation. In addition, because the model's surface data depends partly on the atmosphere, if the atmospheric processes are not fully resolved, over areas with complex topography, the model's surface data may not be sufficiently precise. In these situations, physical postprocessing of the direct model data can add extra details to or improve the quality of the forecast. In this chapter, some physical postprocessing methods are discussed. All these methods assume an unbiased model input. Although these methods are explained in terms of deterministic models, they can be applied to all members of an EPS. Once a range of values is obtained, these values can then be put into a probabilistic context using statistical methods.

2.1 Meteorological diagnostic information

The meteorological diagnostic information presented in this section can be calculated using methods involving prognostic variables (temperature, humidity, wind speed and direction) that are output from operational NWP models.

1) The following diagnostic information concerns the impact of ambient temperature, humidity and wind on human well-being.

   a. **HUMIDEX** (Materson and Richardson, 1979): A dimensionless number representing the level of discomfort people feel that is attributable to humidity during hot days. It is calculated based on the temperature and partial vapor pressure of the ambient air. **(Tier 1)**

   b. **Wind chill index** (Osczevski and Bluestein, 2005): The combined effect of wind and low temperature (below freezing) on how cold the ambient air feels on human skin (see Figure 1). **(Tier 1)**

2) The following diagnostic information concerns the severity of the convective storms that an environment can produce. It is calculated based on the available instability and wind shear data.

   a. **Lifted index** (Galway, 1956) is calculated based on the difference between the temperature of an air parcel lifted from near surface to 500 hPa and the corresponding environmental temperature. **(Tier 1/Tier 2*)**

   b. **Showalter index** (Showalter, 1953) is calculated based on the difference between the temperature of an air parcel lifted from 850 hPa to 500 hPa and the corresponding environmental temperature. **(Tier 1/Tier 2*)**

   c. **Severe Weather Threat (SWEAT) index** (Djurik, 1994) is calculated based on the low to mid-level wind and temperature. **(Tier 1/Tier 2*)**

   d. **Total Total index** (Djurik, 1994) is calculated based on the low to mid-level temperature. **(Tier 1/Tier 2*)**

   e. **George's** K-**index** (George, 1960) is calculated based on the low to mid-level temperature and dewpoint temperature. **(Tier 1/Tier 2*)**

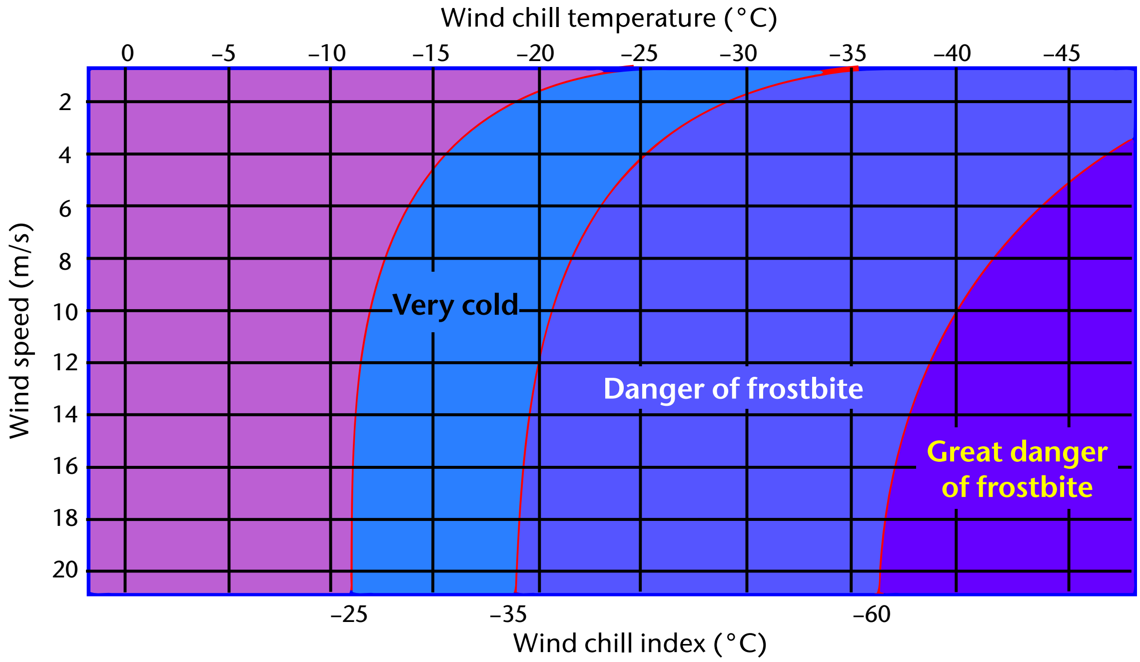

Figure 1. Example of the variation of the wind chill index with respect to temperature and wind speed

Source: https://en.wikipedia.org/wiki/Wind_chill#/media/File:Windchill_effect_en.svg

f. **Storm Relative Helicity index** (Thompson, 2006) is an index that gives information on the low-level environmental wind shear, taking into account both wind speed and wind direction. **(Tier 1/Tier 2*)**

g. **Convective Available Potential Energy (CAPE)** (Markowski and Richardson, 2010) is the theoretical maximum energy available to a convective storm. Conceptually, it is proportional to the kinetic energy that a parcel can gain from its environment as a result of positive buoyancy. This parameter is usually available as a direct NWP output. **(Tier 1/Tier 2*)**

h. **Convective Inhibition (CIN)** (Mapes, 1999) refers to a negative buoyant energy that will inhibit the convection from occurring. Conceptually, it is the opposite of CAPE. The presence of CIN in most cases is at the lower part of the atmosphere. **(Tier 1/Tier 2*)**

* The above diagnostic information is classified as **Tier 1** if it is provided as an NWP model output. If it is not provided as an NWP model output, and instead requires data from multiple vertical levels within the NWP model, it is classified as **Tier 2**.

3) Most existing operational NWP models have condensation schemes that can only produce liquid and solid precipitation as direct model outputs. To obtain mixed-phase precipitation, some diagnostic methods (Bourgouin, 2000; Scheuerer et al., 2016) are used. The calculations for these methods are usually done within the model, but they can be postprocessed outside of the model. They are usually based on model's environmental temperature and assume unbiased input. Diagnostic methods for calculating mixed-phase precipitation are classified as **Tier 1** if they are carried out within an NWP model with precipitation types presented as model outputs. However, if methods for calculating mixed-phase precipitation are carried out outside of an NWP model, they require data on multiple model levels and are classified as **Tier 2**.

2.2 Orographic downscaling (Tier 1)

The surface fields, such as the surface temperature, of a model at a point m1 can be projected over a given region at a corresponding point m2. Depending on the resolution of the model, a direct interpolation may cause a misrepresentation in the continuity of the field once several other neighbouring points are projected. This occurs in particular over areas of complex topography due to height variations between the model topography and the actual region. Postprocessing can be used to downscale the surface temperature, for example, by using a standard lapse rate of L=6.5 °C/km (or the model environmental lapse rate when available) to estimate the temperature difference between points m1 and m2 and then by adding or subtracting the correction to the field at m2. This is done by simply defining the height difference between m1 and m2 and then multiplying this difference by L. The same technique can be used for the neighbouring points. The resulting output is an adapted and continuous model surface temperature projected over the needed region.

Because statistical postprocessing can dramatically improve the quality of forecast products, it has been discussed in the literature for many decades. Early weather forecasts were sometimes adjusted based on a technique known as the "perfect-prog" method (see, for example, Wilks, 2019), which required no training data from the dynamical model. The MOS method began to be used when sufficient dynamical-model training data became available (Glahn and Lowry, 1972). The MOS approach, based on linear or multiple regression, has generally been preferred over the perfect-prog method because the latter is unable to make adjustments to the forecast in situations when the weather forecast is clearly biased.

Figure 2 presents a brief example illustrating the potential benefits of statistical postprocessing. In this example, past numerical forecasts at a certain lead time and the corresponding observations are available for a certain time period in the past. The upper panel shows the temporal evolution of the forecasts generated via the Global Forecast System (GFS) without correction. A clear overall bias is present in the forecasts (estimated from the differences in the dashed lines). Statistical postprocessing aims to adjust the forecast to remove the estimated bias. The lower panel shows the forecasts with corrections based on the classical MOS approach of Glahn and Lowry (1972). This approach will be introduced in Section 3.2.

Slightly more complicated corrections can be made to ensemble forecasts to incorporate corrections for typical overconfidence in raw ensemble forecasts. These corrections fit a probability density function (PDF) using equations to predict a postprocessed mean and spread (standard deviation of the ensemble with respect to its mean). This is illustrated in Figure 3. The

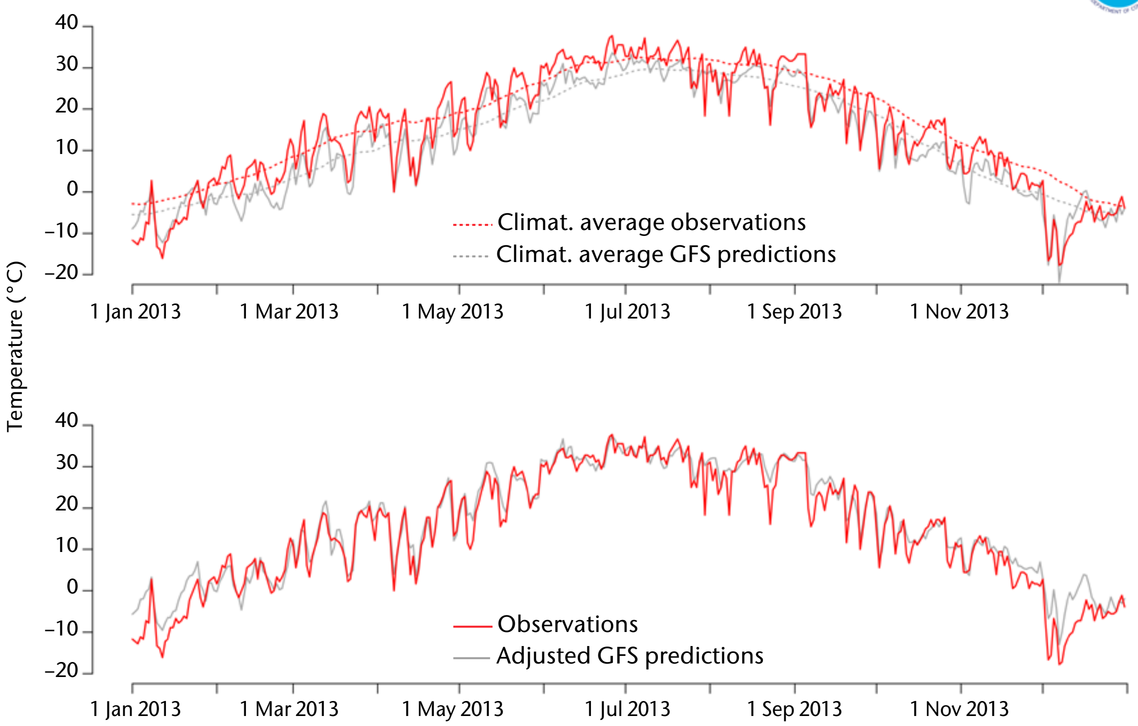

Figure 2. Temporal evolution of predicted (grey) and observed (red) 2-m temperature at Grand Junction Airport (Colorado, United States of America) at +72 h forecast lead time. Upper panel: raw forecast; lower panel: postprocessed forecast. The dashed lines in the upper panel provide smoothed time series to more clearly illustrate the bias.

Source: M. Scheuerer, NOAA/Office of Oceanic and Atmospheric Research/Physical Sciences Laboratory and University of Colorado/Cooperative Institute for Research in Environmental Sciences

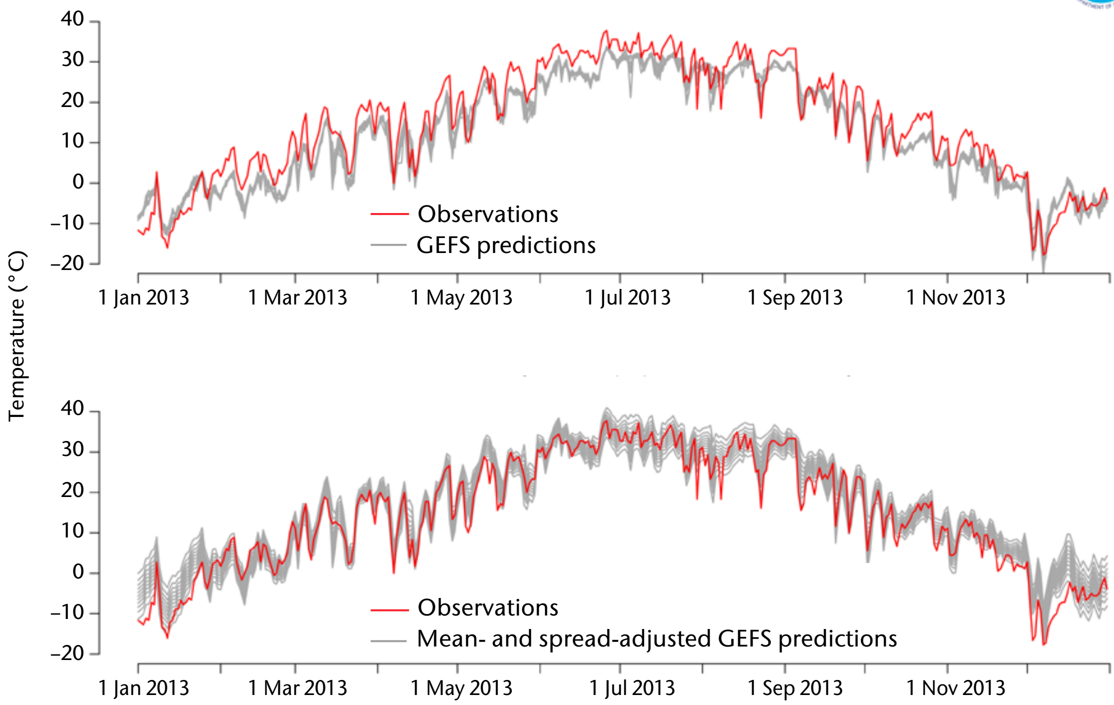

Figure 3. Temporal evolution of predicted (grey) and observed (red) 2-m temperature at Grand Junction Airport (Colorado, USA) at +72 h forecast lead time. Illustration of an ensemble MOS approach to statistically adjusting for both bias in the mean forecast and overconfidence. Top panel: quantiles of the raw ensemble forecast; bottom panel: statistically adjusted quantiles of the ensemble prediction.

Source: M. Scheuerer, NOAA/Office of Oceanic and Atmospheric Research/Physical Sciences Laboratory and University of Colorado/Cooperative Institute for Research in Environmental Sciences

grey area in the upper panel represents the spread of values given by the raw ensemble forecast generated via the Global Ensemble Forecast System (GEFS). If only mean bias is corrected, the ensemble is overconfident and unreliable. However, the observed temperature, represented by the red line, frequently lies outside that spread, so the forecast is overconfident in how accurately it can predict the temperature. The lower panel depicts quantiles of fitted distributions that account for both corrections to the mean and the spread calculated using ensemble MOS (EMOS). Popular methods of mean and spread adjustments are presented in Section 3.3.

Different users may desire different types of postprocessed ensemble outputs. Some may simply want PDFs that are reliable at each observation location or grid point, while others may wish for the postprocessed guidance to still have realistic variability in time and space. For example, the raw forecast for ensemble member 1 may be the windiest member at Los Angeles in the western United States but the least windy member at New York City, and the application (for instance, for airplane flight routing) may need ensembles with realistic variability in space and time but with the biases removed. Methods that process the data "univariately", that is, independently from one grid point to the next, may destroy the underlying rank-order relationships that existed in the original ensemble. The focus of this chapter is on univariate statistical postprocessing, but if the end products are reconstructed ensembles, multiple statistical postprocessing methods may be possible. Multivariate statistical postprocessing will be addressed in Chapter 4.

The remainder of this section presents an overview of several common univariate techniques, starting with simple techniques and progressing to more algorithmically complex ones.

3.1 Deterministic bias correction (Tier 1)

Deterministic bias correction may be suitable for correcting a single-member (deterministic) forecast or adjusting the mean of an ensemble. It is most suited to variables that have Gaussian or near-Gaussian distributions and variables whose systematic errors are relatively large compared to the random errors in the training sample. This method has been used successfully to adjust many forecast variables such as surface temperature, pressure, and winds. Because of its simplicity, this method has been applied widely around weather forecast centres and private sectors for forecast calibration. The Kalman filter-type method (Kalman, 1960), a very simple method of deterministic bias correction, is described below.

1) **Produce a bias-corrected forecast:** The bias-corrected forecast $F_c(t)$ on a particular day of year t and for a given location and forecast lead time is generated by applying a bias estimate $B(t)$ to the current (raw) forecast $F(t)$. These forecasts are computed independently for each lead time and each grid point or station location. When the procedure is initiated, the bias estimate is typically cold-started with a value of 0.0:

$$F_c(t)=F(t)-B(t) \quad 3.1$$

2) **Compute the most recent forecast bias:** When the verifying observation becomes available, the sample forecast bias $b(t)$ is calculated:

$$b(t)=F(t)-O(t) \quad 3.2$$

3) **Update the bias estimate:** The running bias estimate is updated using a weighted combination of the previous day's bias estimate $B(t)$ and the current day's sample bias $b(t)$ using a weight coefficient w:

$$B(t+1)=(1-w)B(t)+wb(t) \quad 3.3$$

An advantage of the Kalman filter-type bias correction method is that it is easy to implement. The system does not need to store or save a long time series of prior forecasts; it only needs to update and save the previous day's estimate $B(t+1)$.

The weight w is the parameter that controls the characteristics of the accumulated bias $B(t)$. An optimal w can be obtained through trial and error. The larger the w, the more the bias estimate reflects the most recent data, which is appropriate if the bias is quite consistent from day to day but exhibits some seasonal dependence. A smaller w correspondingly provides greater weight to samples in the more distant past and is more appropriate if larger samples are needed to quantify the bias and/or the bias is less seasonally dependent. Practically, the optimal w may be some function of the forecast variable, forecast lead time, location or seasonality. Many operational centres use w=0.02 globally for all forecast variables and for all lead times (Cui et al., 2012). When w=0.02, it estimates the bias with the most recent 50–80 days of information (Figure 4). The relative impact of older training samples for other decaying average weights such as 0.01 and 0.05 is also shown in Figure 4. The x-axis is for past days (negative number), and the y-axis is for normalized weight.

An example of the benefit of this simple Kalman filter-type bias correction is shown for northern hemisphere 2-m temperature using a decaying average weight w=0.02 (Figure 5). Verification statistics were calculated over the two-month period ending 27 April 2007 and indicate that mean absolute errors (Wilks, 2019) were reduced by nearly 50% across all lead times.

Like any method, this particular deterministic bias correction has advantages and disadvantages. Its simplicity is a major advantage, as is the lack of any need to store long training data sets. A simple method like this, however, may not provide the magnitude of improvement that is possible with more involved methods, and this method is not applicable for very long training data sets such as multi-decadal reforecasts. It is also not the method that should be used for non-Gaussian-distributed variables such as precipitation amount.

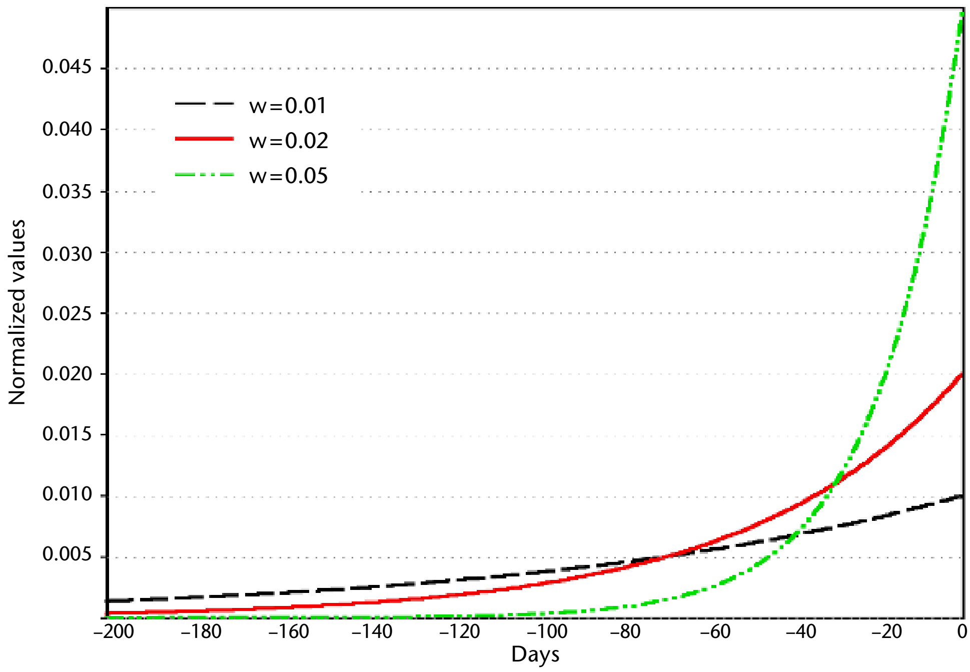

Figure 4. Historical (prior) information (days) used for different decaying average weights (0.01, 0.02 and 0.05). The decaying average weight is shown as a normalized value. The accumulated area (under the curve) for w=0.01, w=0.02 and w=0.05 is equal to 19.

Source: Cui et al., 2012

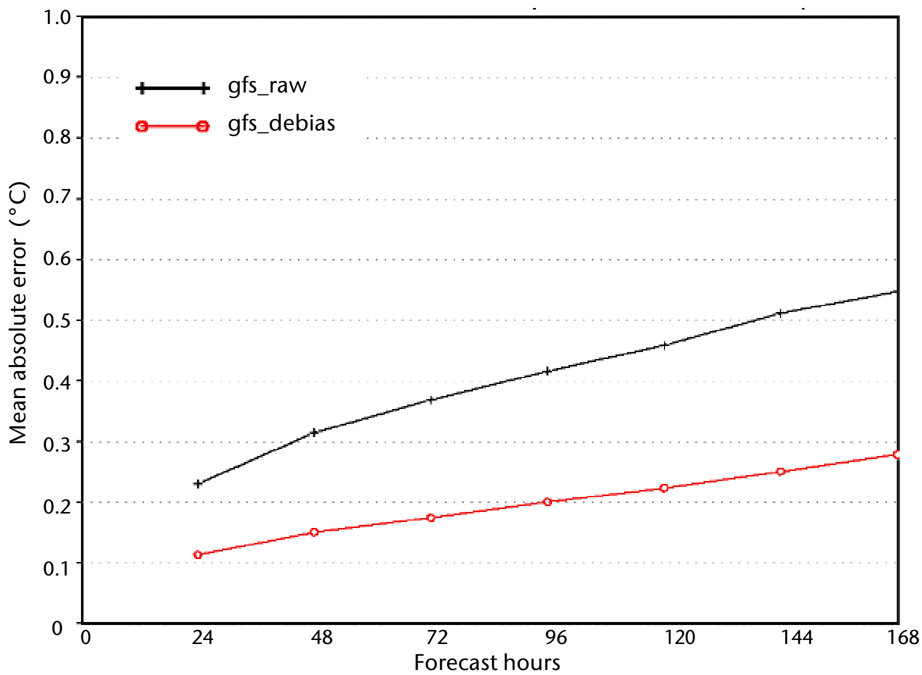

Figure 5. The average absolute error of northern hemisphere 2-m temperature with forecast lead times out to 168 hours (7 days) for the period of 28 February 2007–27 April 2007. National Centers for Environmental Prediction (NCEP) Global Forecast System (GFS) raw forecast errors (gfs_raw) are in black, and bias-corrected forecast errors (gfs_debias) are in red.

Source: Yuejian Zhu, NOAA/Environmental Modeling Center

3.2 Deterministic model output statistics method (Tier 1)

A step up in terms of algorithmic complexity (with respect to the deterministic bias correction method) involves applying regression analysis procedures that permit corrections not only for overall biases but also for some state dependence (for example, different biases for warm versus cold forecast temperatures). The classical MOS correction approach operates on a single deterministic dynamical forecast available at a specific lead time t, applying multiple linear regression:

$$F_c(t) = \alpha + \sum_{i=1}^{N} \beta_i F_i(t) \tag{3.4}$$

$F_c(t)$ is the corrected forecast and $F_i(t)$ is the set of predictors generated by the dynamical model at time t. The parameters α and β_i for i=1, …, N, are estimated by minimizing a cost function, typically the squared difference between past observations and forecasts coming from a training sample, which must be archived in contrast to the decaying average bias correction discussed previously. Often, one of the predictor variables corresponds to the quantity F of interest if it is available as a prognostic variable in the dynamical model.

The forecast performance of the first raw dynamical model outputs was not as high as those of the present day, and the MOS equations sometimes included 10 or more additional predictors (Glahn, 2014), including other dynamical prognostic variables, recent surface observations, climatological values, and some aspects of seasonality or orography (see, for example, Jacks et al., 1990). With improved model performance in recent years, fewer predictors may be needed.

A word of caution about correcting each ensemble member separately

Deterministic forecast errors consist of two components: systematic errors (such as model bias) and random errors. Ensemble forecasts are designed to account for random errors by sampling the synoptically dependent uncertainty in the evolution of weather systems, so the ensemble spread typically increases with increasing forecast lead time. Deterministic forecast corrections such as those described by Equations 3.1 to 3.4 are designed to correct for systematic errors, but the data used to optimize the forecasts also include random errors, which increase with forecast lead time. In applying a MOS procedure to each member of an ensemble with the usual least-square minimization approach to compute regression parameters, minimizing the effect of the random errors results in a "regression to the mean" effect, where forecast corrections adjust each member forecast towards the climatological mean. If such a technique is applied independently to all members of an ensemble, the regression-corrected ensemble members will exhibit a much smaller spread, counteracting the ability of the ensemble to represent random errors.

There are several options to counter this effect and retain the spread in the EPS forecast, including:

(i) Using the perfect-prog approach, where the regression corrections are calculated only at short lead times, when random errors are small, and then applying the same corrections at all forecast lead times. This approach will correct for consistent model biases and errors in representation of local conditions by the model while retaining the ensemble spread but will not correct for any model bias which varies with lead time.

(ii) Using the calculated biases to adjust the ensemble mean forecast at all lead times but then adding the anomalies from the mean of each ensemble member to the corrected ensemble mean, thus retaining the ensemble spread. This approach preserves the original ensemble spread but does not increase it. Increasing the spread is often warranted, however, as raw ensembles are often under-spread, even after correcting for bias.

(iii) Modifying the cost function of the minimization problem, as is done in Vannitsem (2009) and Van Schaeybroeck and Vannitsem (2015), in which constraints are imposed to correct the spread of the ensemble together with its mean. This approach can still be applied to the individual members and provides reliable ensemble forecasts at all lead times.

Alternative methods which calibrate the entire ensemble distribution rather than treating ensemble members individually are outlined in following sections.

Probabilistic forecasts can also be constructed using Equation 3.4, either by estimating a predictive distribution – usually Gaussian – around these corrected forecasts (see, for example, Glahn et al., 2009) or by adjusting each member of an ensemble separately (see, for example, Van Schaeybroeck and Vannitsem, 2011, 2015). This extension of the MOS concept to ensemble forecasts allows both the correction of biases due to model errors as well as a representation of variations in uncertainty based on variations in the ensemble spread.

The classical MOS correction approach, which is based on linear regression, although widely used and an approach which provides important corrections, is difficult to implement when dealing with highly non-Gaussian data sets, such as hourly precipitation. This necessitates the use of alternative approaches. One possibility is to transform the data in order to get an approximate Gaussian distribution in order to implement classical Gaussian approaches (Hemri et al., 2015). A second possibility consists of fitting other predictive distributions, such as logistic or gamma, or even generalized extreme value distributions (see, for example, Wilks, 2018). Combinations of distributions and truncated distributions are also often used (see, for example, Baran and Lerch, 2018, for recent applications).

3.3 Ensemble calibration methods

3.3.1 *Ensemble model output statistics (Tier 2 with site observations; Tier 3 on a grid)*

Another approach to postprocessing involves adjusting the appropriate probability distributions for the ensemble of forecasts. This approach is often referred to as ensemble MOS (EMOS) (Gneiting et al., 2005). One of the most popular EMOS methods is Non-homogeneous Gaussian Regression (NGR), which consists of fitting a Gaussian distribution to the set of ensemble members with a mean that is linearly regressed to the observed training data set, as in the classical MOS method (Equation 3.4), and a variance which depends linearly on the variance of the ensemble (see Wilks, 2018 for more details). NGR has been shown to work very well for variables with a probability distribution close to a Gaussian shape, for instance temperature. For variables such as precipitation or wind, other distributions should be used (see, for example, Wilks, 2018, Scheuerer and Hamill, 2015) and may be more difficult to implement successfully.

An example demonstrating the application of two postprocessing techniques, a member-by-member method called Error-in-Variables MOS (EVMOS), introduced in Vannitsem (2009), and the EMOS-based NGR method, is provided in Figure 6. Both approaches provide an uncertainty band; however, there is a larger variance for NGR due to the presence of representativeness errors. Representativeness errors are not accounted for in the EVMOS method, which only corrects for the impact of model uncertainty at the scale of the model.

In general, the EMOS approach has been used for site forecasts using site observations for training **(Tier 2)**. However, this approach may also be applied to gridded data using trustworthy analysis fields in place of observations **(Tier 3)**. This may be done either by calculating the EMOS coefficients independently at each grid point, by pooling the grid points together to generate a single set of coefficients for the entire domain or by employing an intermediate approach where grid points are pooled over a sub-domain. Calculating the EMOS coefficients independently at each grid point provides optimal local corrections; however, this method requires significant computational resources in order to generate robust EMOS coefficients and an adequate sample size within the training data set for each grid point. The method of generating a single set of EMOS coefficients for the entire domain focuses on correcting a domain-wide bias and will likely require a reduced training data set (compared to the method of calculating the EMOS coefficients independently at each grid point), as all grid points within the domain are used in unison to generate the EMOS coefficients. This method increases the sample size but does not provide location-dependent error corrections. The intermediate method of pooling grid points over a sub-domain produces fewer localized corrections than the method which calculates EMOS coefficients independently for each grid point; however, the corresponding computational requirements and training data set requirements are also reduced. One disadvantage of this method is that there may be discontinuities at the edges of subdomains where biases are estimated from different samples.

EMOS on the grid	*Calculation of coefficients at each grid point*	*Calculation of a single set of coefficients using all grid points*
Advantages	Local corrections	Reduced training data set requirements to compute robust coefficients
Disadvantages	More computationally expensive. A longer training data set likely required to compute robust coefficients.	Only domain-wide corrections possible

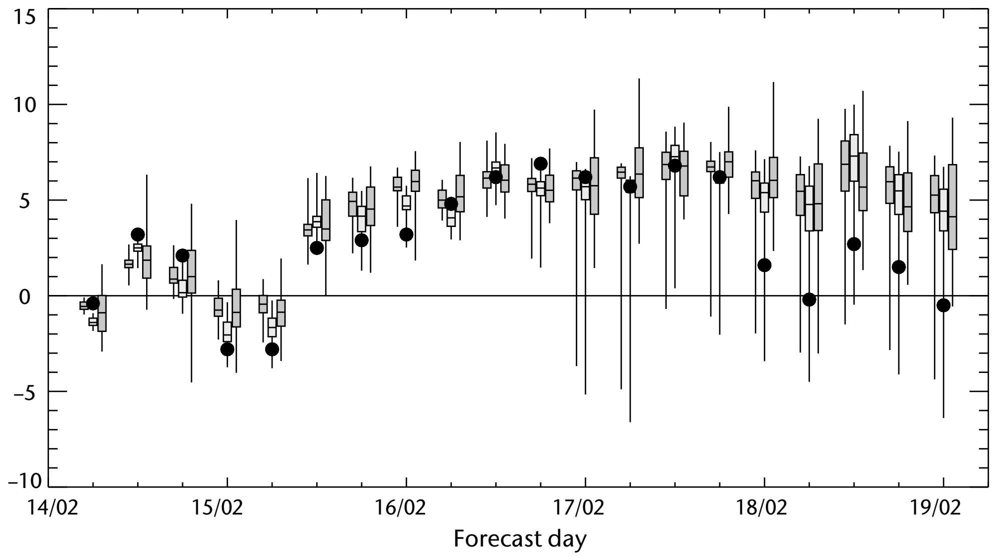

Figure 6. EPS meteograms (EPSgrams) for one specific date (14 February 2009) at Uccle, Brussels (Belgium). The black dots represent the observation. The ensemble distribution is represented by the box-and-whiskers, with 50% of the ensemble lying in the inner box and the remaining members represented by the whiskers. At each lead time, the box-and-whiskers represent, from left to right, the EVMOS-corrected ensemble (dark grey), the raw ensemble (light grey) and the NGR-corrected ensemble (dark grey).

Source: Vannitsem and Hagedorn, 2011

3.3.2 ***Bayesian model averaging (Tier 2)***

Another important postprocessing approach involves combining a large set of probability distributions centred around or associated with a specific ensemble member. The most widely used method of postprocessing in this manner finds its roots in the Bayesian framework (Raftery et al., 2005), although it is not fully Bayesian (see Wilks, 2018 for a discussion on this topic). This method is known as Bayesian model averaging (BMA) and consists of constructing a corrected probability distribution as

$$f\left(F_c\left(t\right)\right) = \sum_{i=1}^{N} w_i f_i\left(F_i\left(t\right)\right) \qquad 3.5$$

where w_i represents the weights given to each distribution $f_i(F_i(t))$ associated with each ensemble member i=1, …, N. These weights and the distribution parameters are fitted based on a past training data set. BMA yields a continuous predictive distribution for the forecast variable F. However, rather than imposing a particular parametric form, BMA predictive distributions are mixture distributions, or weighted sums of N-component probability distributions, each centred at the corrected value of one of the m ensemble members being postprocessed. This approach has been used widely for both exchangeable members (where all ensemble members are equally likely by definition) and non-exchangeable members (see, for example, Sloughter et al., 2010; Baran, 2014). Other postprocessing methods corresponding to this approach include ensemble dressing, which was proposed prior to the development of the BMA method (Roulston and Smith, 2003), and other Bayesian alternatives (see, for example, Marty et al., 2015).

There are disadvantages and reasons to be cautious of the BMA approach. These are discussed in such publications as Hodyss et al. (2016) and in the references therein.

3.4 **Quantile mapping (Tier 2)**

For some variables such as precipitation, bias correction may depend on the precipitation amount and may involve over-forecasts at low precipitation amounts and under-forecasts at high precipitation amounts. Techniques that are flexible enough to account for state-dependent biases are thus extremely useful. One approach to addressing state-dependent biases is quantile regression (Bremnes, 2004), which can also be of much interest when the user does not want to specify a distribution.

A commonly applied procedure, known as quantile mapping, is illustrated in Figure 7. Quantile mapping can be applied to a deterministic forecast or individually to members of an ensemble. In Figure 7, two cumulative probability distributions are represented, one for the forecast and a second for the observations, which are obtained via a large ensemble of past forecasts. The red arrows show how the quantile mapping operates: given a forecast of the current day's precipitation amount, the quantile associated with this amount is identified, and the forecast amount is replaced with the observed or analysed amount associated with the same quantile in the cumulative distribution.

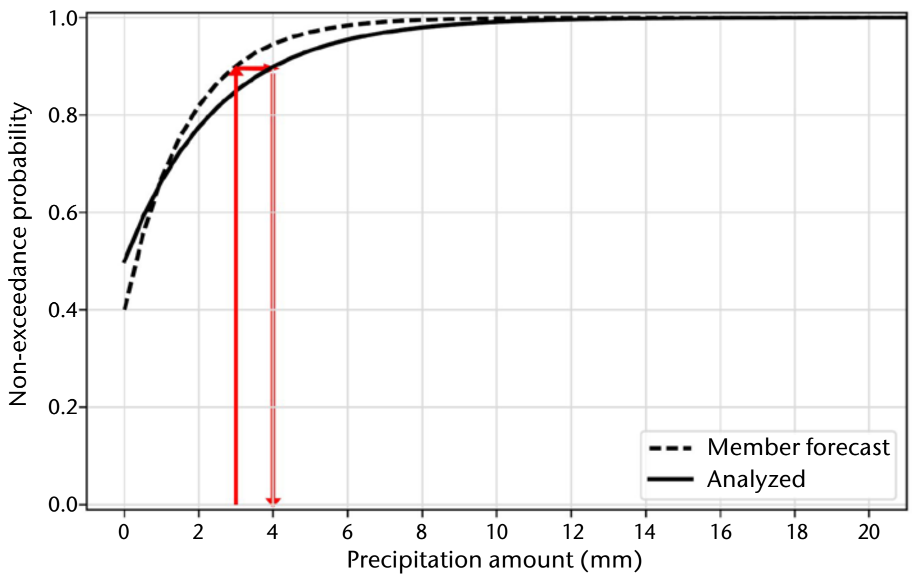

Figure 7. Illustration of the deterministic quantile mapping method applied to ensemble members. Member forecast cumulative distributions and analysed cumulative distributions are used to adjust the raw forecast value to the analysed value associated with the same cumulative probability. Red arrows denote the mapping process from forecasts to the associated quantile determined from training data and back to the analysed quantity associated with this quantile.

Source: T. Hamill, from Chapter 7 in Vannitsem et al., 2018

Quantile mapping is an intuitive and appealing technique in concept. In practice, the differences between the mappings of forecast and analysed/observed distributions may be very large and prone to sampling variability unless large samples are used to generate the underlying cumulative distribution functions (CDFs). In practical applications, data from supplemental locations (other locations with presumed similar forecast characteristics) may be used to populate the CDFs (Hamill et al., 2017), and/or more conservative mapping approximations may be used in the tail of the distribution (Hamill and Scheuerer, 2018) to prevent unrealistically large mappings.

3.5 **Machine learning**

Statistical learning or machine learning methods are used in a wide range of postprocessing approaches in order to extract some key representative features from a set of raw data. These representative features can then be used to infer new outcomes based on new information entering the machine learning algorithm. Machine learning methods are very successful at recognizing patterns in image processing (see, for example, Goodfellow et al., 2016).

Deep learning is a particularly effective machine learning method in which the original complex pattern is decomposed into much simpler patterns. Paradigmatic examples are the multi-layer perceptron and the multi-layer neural network that allow the complex link between an input and an output to be decomposed into a multitude of simpler connections between nodes, at which linear or nonlinear functions can transform their own input (see, for example, Goodfellow et al., 2016). This type of method was used quite early in the atmospheric sciences (compared to other fields) for both statistical forecasting (Hsieh and Tang, 1998) and forecast postprocessing (Marzban, 2003; Casaioli et al., 2003) purposes. Early learning approaches used simple forms of neural networks, usually with a single layer between the input and the output. Since then, considerable progress has been made in increasing the types of algorithms (Goodfellow et al., 2016) and also the efficiency of the computing.

Simple machine learning techniques, such as employing a single-layer neural network using the standard software routines available in Python libraries, for example, may provide a quick and effective **Tier 2**-level alternative to some of the statistical techniques, such as MOS, described above. However, care should be taken to ensure that sufficient training data is used to cover a wide range of situations and to keep the neural network simple; otherwise, there is a risk of over-fitting to a small data sample, leading to poor and misleading behaviour with future forecasts.

Ensemble learning methods are techniques used for classification and regression based on building decision trees in a training data set. As these are known to be highly dependent on the sample used and prone to over-fitting, randomization is used through bootstrapping and random predictor selection. The result of this randomization is known as a random forest (RF). Once a new realization of predictors is presented to the trees, the split is followed until a leaf is reached. The average situation associated with this leaf from the training sample is the forecast, which is then averaged through all the trees of the random forest. A further development to get distributions instead of the conditional mean has also been proposed and is known as quantile regression forests (QRFs) (Taillardat et al., 2016). An example of the application of the QRF method is displayed in Figure 8.

Machine learning techniques will generally be **Tier 2** or **Tier 3** due to the requirements for managing large quantities of data for training and testing. Whatever machine learning technique is chosen, a key requirement for success is careful attention to the quality control of the data used, which can often be the most time-consuming part of the work.

A comparison between neural networks, EMOS and BMA was performed in Rasp and Lerch (2018) and demonstrated the usefulness of adopting machine learning approaches. These approaches can, however, be very demanding in terms of training data.

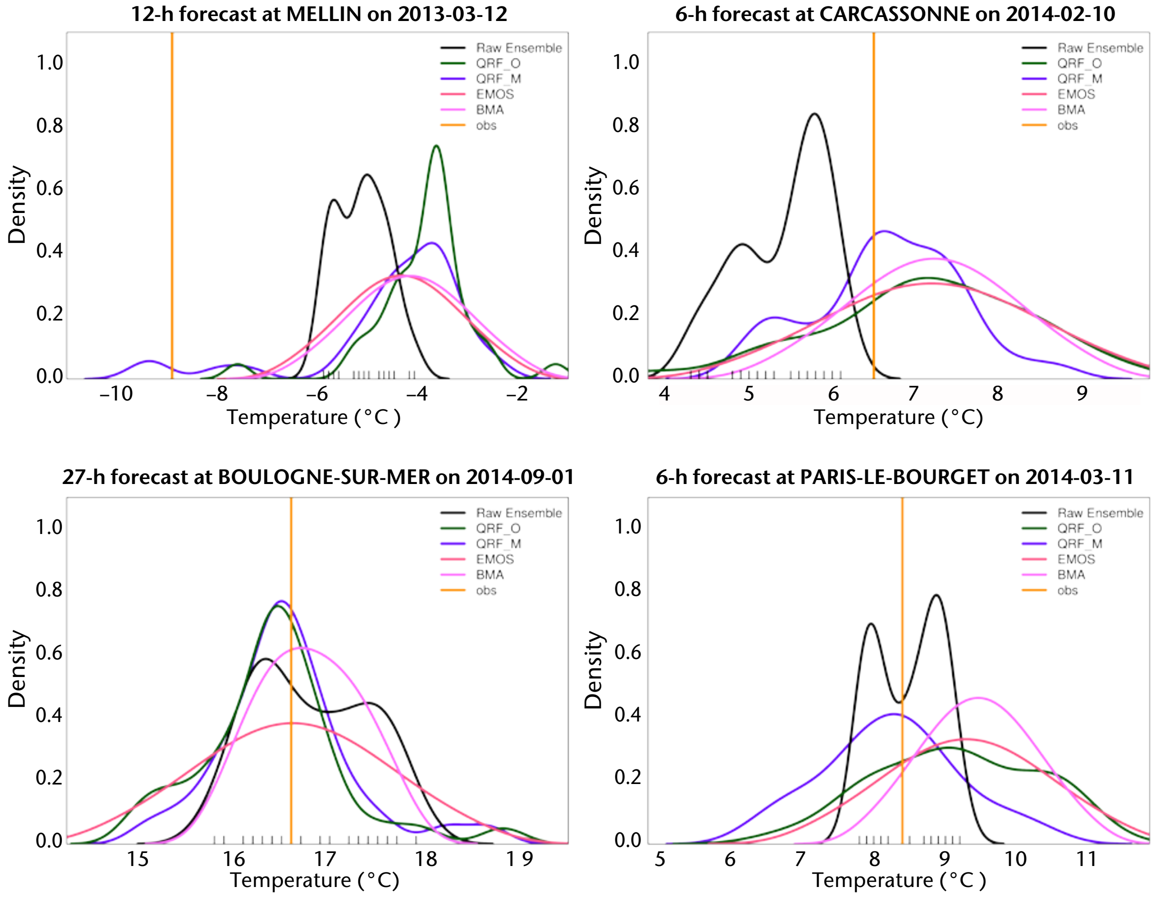

Figure 8. Some forecasts for different meteorological situations where the QRF_M technique, a multivariate variant of QRF, is useful for forecasters. (For comparison, the QRF_O technique is the univariate variant of QRF.) Top left: The QRF_M technique proposes cooler scenarios. Top right: The bimodality of the raw ensemble is preserved. Bottom left: Bimodality is still conserved, but one mode is preferred over the other. Bottom right: The QRF_M technique proposes a unimodal PDF contrary to the raw ensemble. The small segments on the x axis represent the 35 raw members. Several members are associated with the same temperature.

Source: M. Taillardat, Météo-France/COMPAS – CNRM UMR 3589

3.6 Analogue methods (Tier 3)

Analogues (or analogs) were first introduced in the 1960s to evaluate the predictability of large-scale atmospheric flows (Lorenz, 1969). The idea was to find similar atmospheric situations in a catalogue of three-dimensional pressure fields that were recorded during a certain period and to investigate the error amplification between forecasts starting from these analogues. At the time of the work of Lorenz, only five years of daily data were available to him. With these large-scale fields, Lorenz did not find very good analogues and hence could only reach quite poor estimates of the predictability of large-scale atmospheric fields. At present, the size of the catalogues is much larger, with longer periods of coverage and higher frequencies of recording. It is therefore currently possible to obtain better analogues even at very large scales; however, the possibility of finding "good" analogues that would allow high quality forecasts to be generated is still very probably quite unrealistic (see, for example, Nicolis, 1998, and the references therein).

The analogue approach is nevertheless very useful, as illustrated by the multiple applications that have been made of this empirical approach to assemble similar dynamical fields. One prominent example is the use of analogue methods for downscaling. For example, it may be that there are two different data sets available, one at coarse resolution and another at high resolution. Furthermore, there may be a map relating some of the coarse resolution fields to the high-resolution ones. Usually, the coarse resolution catalogue is much larger than the high-resolution catalogue. In this instance, the analogue approach seeks to find coarse resolution analogues to a specific target situation for which there is no high-resolution correspondence and to associate to the target the high-resolution field(s) of the analogue situation(s) (see, for instance, Maraun and Widmann (2018) for a general description and more specific references on this topic).

For statistical corrections, the analogue approach has been shown to produce results comparable to those produced with more traditional approaches (Hamill and Whitaker, 2006; Delle Monache et al., 2011, 2013; Nagarajan et al., 2015). The various steps in this approach are nicely illustrated in Figure 9. In this example, the first column contains the targeted forecast (here, for precipitation), which could be a single deterministic forecast or a specific moment of an ensemble forecast, for instance the ensemble mean. In the catalogue of past forecasts, four analogues are selected based on a specific distance (second column). Once this selection is made, the corresponding observations (reanalyses or point observations) are selected (third column) and used to make probabilistic forecasts (fourth column). The verifying observation is provided in the last column. One very interesting advantage of this approach is that higher-resolution observations can be used, providing a statistical downscaling. Another

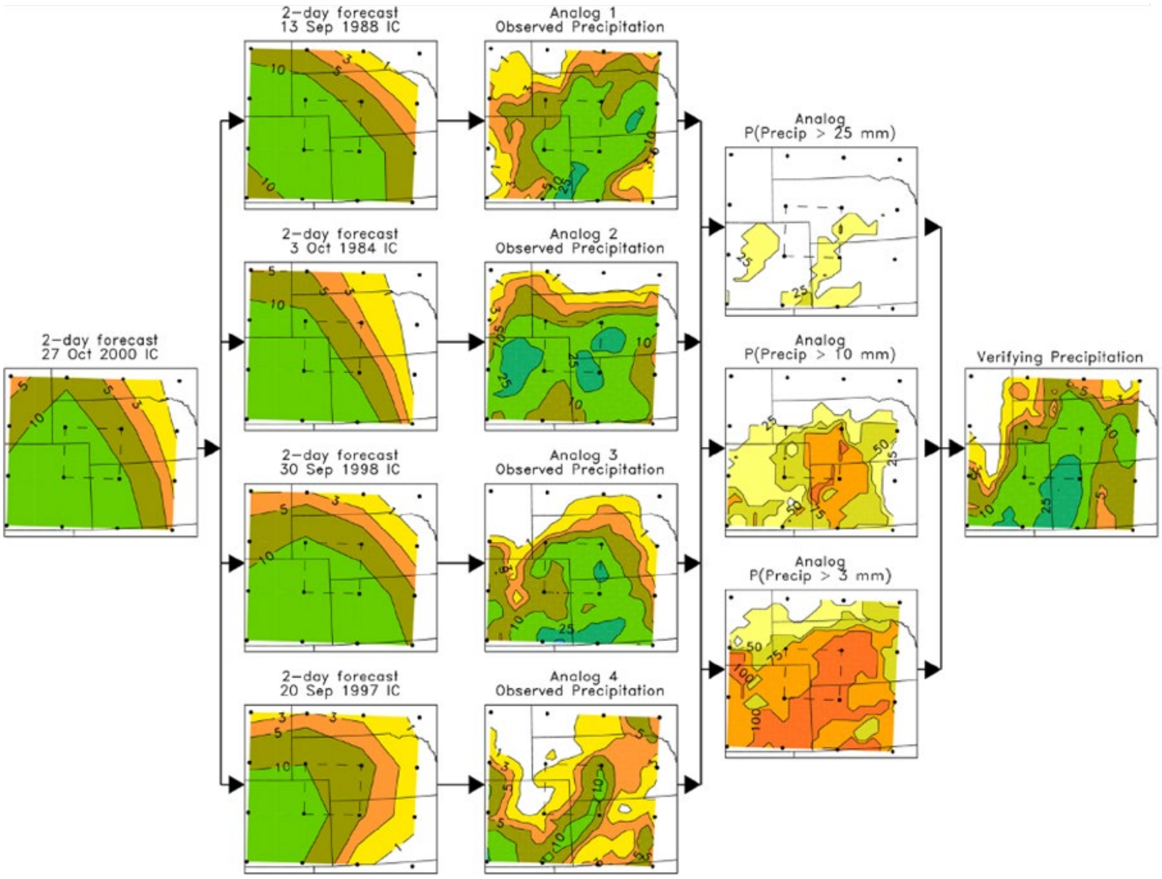

Figure 9. Illustration of the basic analogue technique for a two-day forecast
The coarse-mesh ensemble mean precipitation forecast is shown in the first column, defined at the 16 dots and contoured in the graphic. Analogues and probability forecasts are desired for the dashed box in the middle. The four closest matching two-day ensemble mean forecasts are shown in the second column, and the higher-resolution analysed weather on those dates is shown in the third column. Probabilistic forecasts formed from the analysed analogues are shown in the fourth column for 3-mm, 10-mm, and 25-mm thresholds, and the analysed data are shown in the far right column. In practice, many more than four analogues are used to estimate probabilities.

Source: Hamill and Whitaker, 2006

advantage is its generality: the procedure can be applied similarly to many forecast variables (wind, temperature, precipitation, and so forth). A disadvantage is that it typically requires larger training samples to produce results as accurate as those obtained with other methods.

3.7 Statistical downscaling using high-resolution observations/analyses (Tier 3)

A high-resolution precipitation analysis, such as that produced by the Integrated Nowcasting through Comprehensive Analysis system (INCA) (Haiden et al., 2011) at the Austrian Central Institute of Meteorology and Geodynamics (ZAMG), contains information such as radar or satellite data applicable to areas in between observation sites. Such an analysis can provide observations to be used for statistical postprocessing in order to produce spatial forecasts. However, with a spatial resolution of 1 km and over 250 000 grid points, individually postprocessing every grid point can lead to spatial inconsistencies and is computationally expensive. Dabernig et al. (2017) introduced a method which allows all stations to be postprocessed simultaneously. This method was adapted in Dabernig et al. (2019) to enable all grid points of an analysis to be postprocessed simultaneously.

The basic idea in Dabernig et al. (2017) involves the use of standardized anomalies of observations and NWP forecasts. These standardized anomalies do not contain any station-specific characteristics, which allows all stations/grid points to be forecasted with one single non-homogeneous regression (NHR) simultaneously. This method is called standardized anomaly model output statistics (SAMOS). To standardize observations and forecasts, a climatological mean is subtracted from the daily forecast and divided by a climatological spread, as shown in the following equations:

$$y^* = \frac{y - \mu_y}{\sigma_y} \qquad 3.6$$

$$m^* = \text{mean}\left(\frac{ens - \mu_{ens}}{\sigma_{ens}}\right) \qquad 3.7$$

$$s^* = \text{std dev}\left(\frac{ens - \mu_{ens}}{\sigma_{ens}}\right) \qquad 3.8$$

with μ as the climatological mean, σ as the climatological spread, y as the observation, ens as the ensemble forecast and y^*, m^* and s^* as the standardized anomalies of observation, ensemble mean and ensemble spread, respectively.

A forecast example for temperature is shown in Figure 10. The 1-km resolution in INCA displays clear temperature differences between valleys and mountains, whereas the European Centre for Medium-Range Weather Forecasts (ECMWF) ensemble mean forecasts, with a resolution of ~18 km, are not able to reproduce any valleys. SAMOS is able to reintroduce the

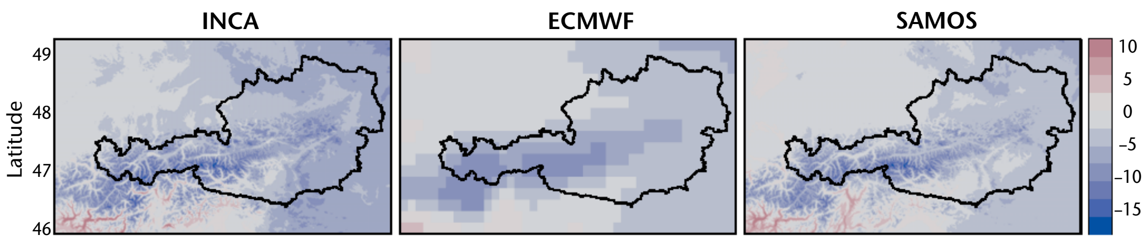

Figure 10. SAMOS temperature forecast example from 7 January 2017 and lead time +36 h, with the analysis on the left, the raw ECMWF ensemble mean forecast in the middle and the postprocessed SAMOS forecast on the right

Source: Dabernig et al., 2019

complex structure of alpine valleys into the larger forecast pattern of ECMWF. Further details, other variables and verification compared to a station-based method can be found in Dabernig et al. (2019).

To summarize, SAMOS allows all grid points in a 1-km resolved analysis to be postprocessed using a parsimonious method, preserves the spatial pattern and is similar in skill to a point-wise NHR.

3.8 Spatial methods – neighbourhood processing (Tier 2)

Most postprocessing methods are based on an assumption that the NWP model or ensemble can approximately predict the spatial location of weather events. Errors in location or timing of such events are common but are particularly severe for very high-resolution convection-allowing models and ensembles in which convective storms can be explicitly represented. While the models may represent convective storms realistically within a region, the precise location and timing of storms is very unlikely to be predictable. Neighbourhood processing methods have been developed specifically to address this problem. The assumption in a neighbourhood approach is that a storm prediction centred on one model grid cell is equally likely to occur at a number of grid cells within a neighbourhood around that location. This is illustrated in Figure 11 for an event (for example, rain rate exceeding 1 mm/hr) which occurs at a number of grid cells that are highlighted in grey. The probability of rainfall occurring at a grid cell is calculated as the proportion of cells within the neighbourhood where the event occurs – in Figure 11, a neighbourhood of 5 x 5 cells is used. Use of a larger neighbourhood will smooth the spatial uncertainty more, but use of an excessively large neighbourhood risks losing the resolution of predictable detail that is available from the models.

Neighbourhood methods can be used to generate a probability estimate from a single deterministic model forecast where there is spatial uncertainty but can also be used for every member of an ensemble forecast. With an ensemble, the neighbourhood method will smooth unrealistic spatial variability in the probabilities from the raw ensemble by, in effect, generating a much larger ensemble size with which to estimate the probability at a grid location. The ensemble members account for uncertainty due to the larger-scale synoptic evolution of the forecast, while the neighbourhood takes account of local spatial uncertainty. In practice, to achieve reasonably smooth probability fields, a much smaller neighbourhood is generally required with an ensemble than with a deterministic model, allowing better retention of spatial resolution in the forecast.

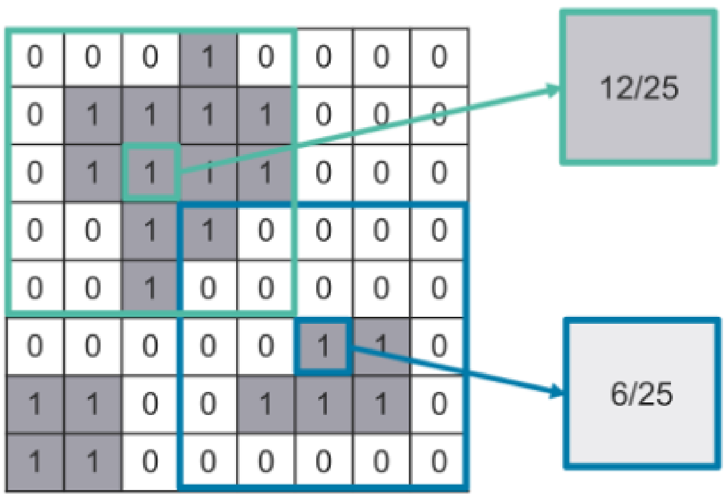

Figure 11. Illustration of the use of a neighbourhood of grid cells, in this case, a neighbourhood of 5 x 5 cells, to estimate the probability of an event occurring in the central cell when there is spatial uncertainty in the prediction

Source: F. Rust, N. Roberts and K. Mylne, Met Office UK © British Crown copyright

The basic neighbourhood method is most applicable for precipitation from convection-allowing models and EPSs, but some additional enhancements may be required for other applications. It may be desirable to use the same size neighbourhood for all variables for consistency, but in many cases where elements are less spatially uncertain, it may be more appropriate to use smaller neighbourhoods for some variables. A simple neighbourhood as described above may not be appropriate where the weather element is sensitive to local geographic effects such as topographic height or coastal effects, and some enhanced methods are given below:

- Topographic neighbourhood: For temperature or wind speed, for example, where the value varies significantly with orographic height, a topographic neighbourhood can be used to select only those neighbourhood grid cells with a similar orographic height. An alternative approach might be to make orographic adjustments of values to the height of the central cell. For some elements, such as visibility (or fog) occurring in valleys or over hilltops, this method may be important to retain model-resolved detail.

- Orographic enhancement: Where precipitation occurs over a region with significant orographic variation, some orographic adjustment of precipitation rates may be required to adjust values from one grid cell for use at a neighbouring location.

Software and documentation for neighbourhood processing, including topographic neighbourhoods, are available at https://github.com/metoppv/improver.

3.9 Extreme events

Reliable estimation of the probability of extreme events is important for many applications such as the issuing of risk-based severe weather warnings. However, the calibration or bias correction of extreme event forecasts presents particular challenges for statistical postprocessing because extreme events are (by definition) rare in the data samples. For example, a simple bias correction of a temperature will reflect the mean error of a model across a wide range of weather types, but it is likely that when the model forecasts an exceptionally high (or low) temperature, its error characteristics may be quite different, and the bias “correction” applied by postprocessing may be inappropriate, possibly even in the wrong direction, making the forecast worse rather than better. There are a number of methods which may be used to better account for extreme forecasts in statistical postprocessing. Caution should also be exercised in assessing the quality of any correction for extreme events since the sample size in any verification is likely to be small and may not be representative of future extreme events.

- The use of more complex state-dependent corrections, for example, the development of a regression line (or curve) to estimate the bias as a function of the forecast value, may better estimate the expected bias for an extreme forecast. One approach may be to use an extreme value distribution to estimate the error for an extreme forecast. However, care must be taken not to over-fit estimates to very small samples of extreme events which may not be representative of future errors. **(Tier 1)**

- Various methods may be used to increase the sample size of statistical training data for extreme events:

  – Reforecast data sets are available for some major EPSs (see, for example, Vitart et al., 2019). These data sets provide re-runs of the forecast system for many previous seasons. Combined with the corresponding observations, reforecasts can be used to train a calibration over a much longer period with consistent forecast performance, better capturing extreme events (Hamill et al., 2004; Hamill, 2012). (This approach requires access to long archives of quality-controlled observations and the management of very large quantities of data.) **(Tier 3)**

  – Another approach to increasing the data sample, which may be applied for training a calibration for a location without access to reforecast data sets, is to identify other supplemental locations with similar characteristics and therefore similar forecast error characteristics. Data for these supplemental locations can be used to provide a much larger sample for calibration at the location of interest (Hamill et al., 2017). **(Tier 3 due to complex data handling)**

- A simple approach to avoiding the use of inappropriate "corrections" where there is a concern that the method used may not be appropriate for extremes is to apply a weighting to the correction which reduces towards zero as the forecast value becomes more extreme, thus reverting the forecast back to the raw model forecast in extreme situations. One example of this approach is given by Equation 6 in Hamill and Scheuerer (2018). **(Tier 1)**

One of the challenges for forecasting extreme events is the limited extent to which NWP models may be able to fully represent meteorological extremes, or of the ensemble to capture the extreme event within the ensemble spread (sampling). An effective way to compensate for this is to consider forecasts in terms of the predicted anomaly from the model climate rather than in absolute terms. By relating the model climate to the real observed climate using a technique such as quantile mapping (see Section 3.4), the predicted anomaly may be related to expected extremity in the real climate. For this purpose, the model climate is best defined using a set of reforecasts (see Section 7.4), where available, or otherwise from as many recent forecasts as can be accumulated. Anomaly forecasts may be expressed in different ways, but examples may include:

- Probability of exceeding the 90th or 99th centile of climatology;

- Probability of exceeding two standard deviations above or below the climatological mean.

Where the model and observed climatological data are readily available, these methods are classified as **Tier 2**, but if these data have to be generated, the methods are **Tier 3**.

Some advanced postprocessing methods have been developed at major centres specifically to address extreme event prediction from global ensembles. Two examples, the Extreme Forecast Index and Point rainfall, are given below.

3.9.1 ***Extreme Forecast Index (Tier 3)***

The Extreme Forecast Index (EFI) (Lalaurette, 2003; Zsoter, 2006) is an example of a postprocessed product that is specifically designed to provide guidance for anomalous, extreme or severe weather events (for example, heavy precipitation, strong winds, extreme temperatures, or unusually high ocean waves). EFI compares the ensemble forecast probability distribution to the distribution of the model climate for the chosen weather variable, location, time of year and forecast lead time. The underlying assumption is that, if a forecast is anomalous or extreme with respect to the model climate, the real weather is also likely to be anomalous or extreme compared to the real climate. EFI takes values between +1 (all members are above the climate maximum) and -1 (all members are below the climate minimum). The closer EFI is to +1 or -1, the more likely it is for extreme weather to occur. Complementary products that focus on the magnitude of the extreme, such as the Shift of Tails (SOT), can also be produced.

The model climate that is needed to compute EFI is usually generated from reforecasts provided by the producing centre. More information about EFI and SOT can be found in the ECMWF Forecast User Guide.

3.9.2 ***Point rainfall (Tier 3)***

The model output fields from an ensemble forecast are grid box average values. For precipitation, for example, they represent the average rainfall over the grid box, which can represent an area of 20 x 20 km to 100 x 100 km for global ensemble model forecasts. The observed rainfall totals at an individual location (such as a particular rain gauge) can be quite different from these area-average values. For extreme rainfall in particular, it is important to take account of this difference, and postprocessing methods have been developed to address this issue.

ECMWF point rainfall probability (ecPoint) (Pillosu and Hewson, 2017) is an example of such a postprocessed product. A comparison of raw ensemble and point rainfall charts is presented in Figure 12. Postprocessed products add value to existing products by accounting for the different degrees of sub-grid variability and grid-scale bias that exist in different weather situations. For calibration, ecPoint uses short-range control run 12-hour rainfall forecasts covering one year (the "training period") which are individually compared with rainfall observations for the same times within the respective grid boxes. This procedure involves segregation according to grid box weather types, each of which has different sub-grid variability structures and/or different associated bias corrections. The 12-hour point rainfall system introduced into operation in April 2019 incorporated 214 such types. The type definitions are currently based on the following parameters: convective rainfall fraction, total 12-hour precipitation forecast, 700 hPa wind speed, Convective Available Potential Energy (CAPE) and 24-hour clear-sky solar radiation. For more information on the point rainfall product and the postprocessing methodology used to generate it, see:

- ECMWF Forecast User Guide (https://confluence.ecmwf.int/display/FUG/Point+Rainfall)
- https://www.ecmwf.int/en/newsletter/159/news/new-point-rainfall-products-eccharts.

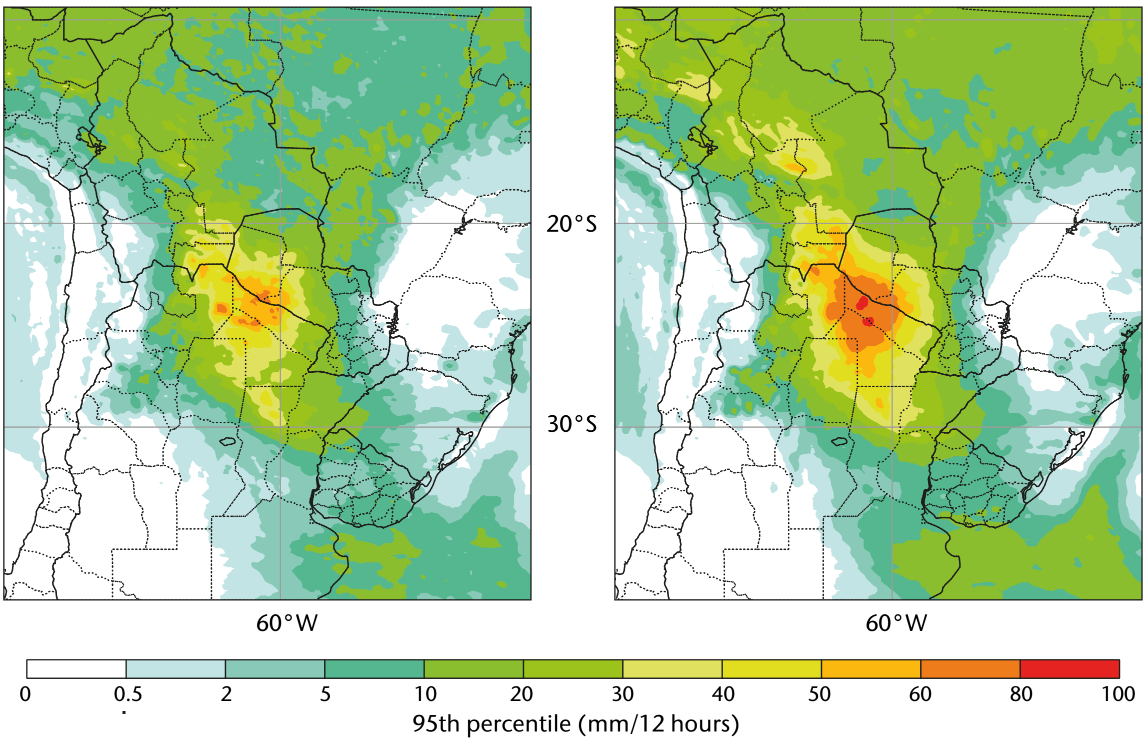

Figure 12. Comparison of raw ensemble and point rainfall charts. The charts show the 95th percentile for the raw ECMWF ensemble forecast (left) and for the corresponding ecPoint rainfall forecast (right) for 0600 to 1800 UTC on 2 April 2019 (T+90 to T+102). Although it is occasionally not the case, in this particular example, the areas where there is a 5% chance of large accumulations, for example 60 mm or more in 12 hours in the Argentina–Paraguay border region, are more extensive in the point rainfall product than in the raw ensemble.

Source: https://www.ecmwf.int/en/newsletter/159/news/new-point-rainfall-products-eccharts

CHAPTER 4. MULTIVARIATE POSTPROCESSING

In Chapter 3, several approaches to the statistical postprocessing of univariate distributions were briefly presented. It may also be of interest to Members to utilize forecasting fields that display either spatio-temporal coherence or inter-variable dependence, or both (Schefzik and Möller, 2018).

Two main routes to preserve these properties are illustrated in Figure 13. Starting with the raw ensemble (top left box), one route consists of applying linear statistical postprocessing to each member of the ensemble (arrow from top left to bottom left), an approach currently known as member-by-member postprocessing (MBMP). This approach usually preserves the correlation structure of the forecasts (see, for example, Van Schaeybroeck and Vannitsem, 2015). A second route, by which the univariate raw PDF built on the raw ensemble is statistically postprocessed, is displayed from top left to top right and from top right to bottom right. At this stage, the spatio-temporal correlation could be heavily modified, and a final stage is therefore needed in order to recover the spatio-temporal coherence of the fields. The first section of this chapter provides more discussion of methods appropriate to multivariate postprocessing.

An entirely different approach involves combining multiple sources of forecasts. This approach is usually referred to as blending. One method for blending is presented in Section 4.2.

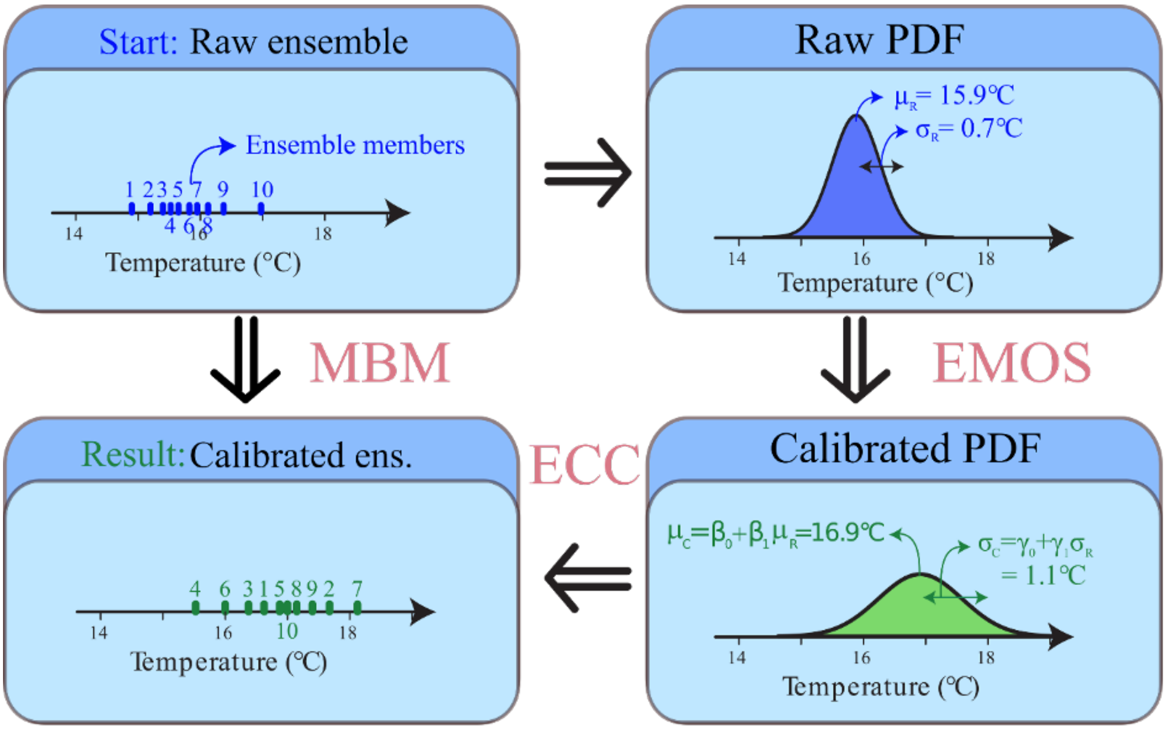

Figure 13. Two routes for postprocessing ensemble forecasts. Top left to bottom left: Transformation of single ensemble members. Top left to top right: Construction of the raw PDF, followed by (top right to bottom right) calibration of the PDF and, if necessary (bottom right to bottom left), a reordering of the members to reassign spatial or temporal consistency. The acronyms MBMP (member-by-member postprocessing), EMOS (ensemble model output statistics) and ECC (Ensemble Copula Coupling) refer to specific techniques that can be used along these two routes. The symbols μ and σ represent the mean and the standard deviation of the distributions, respectively. The subscripts R and C refer to the raw and corrected ensembles, respectively.

Source: B. Van Schaeybroeck, Royal Meteorological Institute of Belgium

4.1 Ensemble copula coupling (Tier 3)

Two different strategies to preserve coherence in distribution-based models have been proposed in the literature. One strategy, which is useful when dealing with a small number of predictands, is to encode this information in a specific parametric multivariate distribution (see, for example, Schuhen et al., 2012). A second strategy, which is particularly suitable when dealing with a large number of predictands, is to adopt a non-parametric approach to preserving coherence. Ensemble copula coupling (ECC) is a prominent example of this second strategy; the Schaake shuffle is another important example (Schefzik, 2017).

ECC (and also the Schaake shuffle) is based on the reordering of samples produced by postprocessed univariate predictive distributions depending on a dependence template linking the different predictands.

Figure 14 illustrates the impact of using ECC in order to preserve the spatial correlation between two stations in Germany: Berlin and Hamburg. The three panels on the left show the raw ensemble forecast issued by ECMWF and valid at 0000 UTC on 13 January 2011. The red dots represent the 51 joint events of the ensemble forecast, and histograms are provided for each station. A clear joint dependence structure is visible between the two stations. In the middle three panels, postprocessing using BMA is performed separately at each station, and the original correlation structure is lost. When ECC is applied to the data after BMA, a new dependence structure appears, as illustrated in the three panels on the right.

Note that ECC preserves correlations that were already present in the original spatial fields. If large errors exist in these dependences, these errors will be propagated to the postprocessed fields.

4.2 Blending of nowcasts and short-lead NWP forecasts (Tier 3)

A typical application of postprocessing at short leads is the production of a high-resolution rainfall forecast for leads of 1–2 hours. Quantitative precipitation nowcasts are mainly based on the advection of rainfall fields observed by meteorological radar. This type of forecasting, called Lagrangian extrapolation, is characterized by high initial skill that rapidly decreases with forecast lead time, as shown by Golding (1998). The two main sources of error in Lagrangian persistence forecasts are the growth and decay of the precipitation and the temporal evolution of the advection field (Tsonis and Austin, 1981; Radhakrishna et al., 2012), and Germann

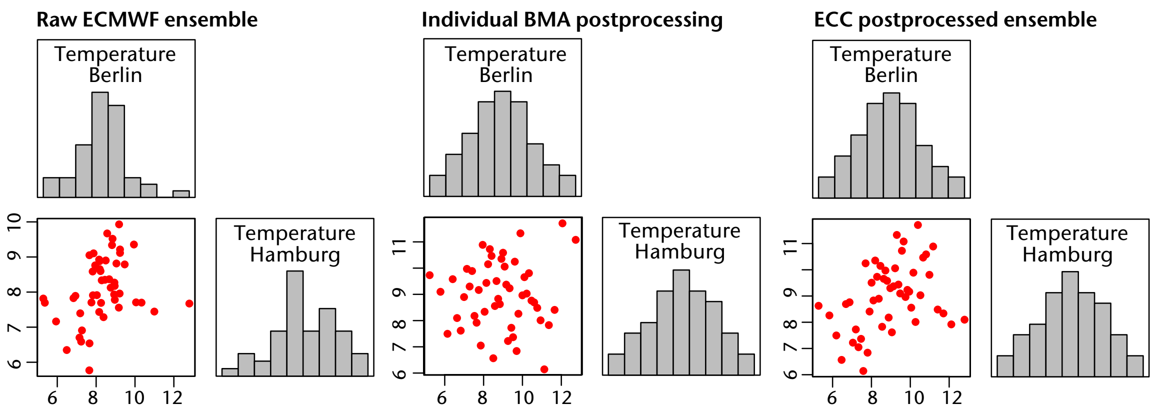

Figure 14. Left: Raw ensemble forecast of 2-m temperature issued for Berlin and Hamburg. The joint events are represented by the red dots, and the histograms represent the univariate distributions. Middle: Univariate postprocessing by BMA is provided for both stations. Right: Ensemble obtained after applying ECC. Ensemble forecast at 24-h lead time valid at 0000 UTC on 13 January 2011.

Source: Roman Schefzik, Medical faculty Mannheim of Heidelberg University, Germany and Zawadzki (2006) showed that the importance of these two factors is case-dependent. Extrapolation-based forecasts are superior to NWP forecasts for lead times of on average three hours depending on the meteorological situation (Simonin et al., 2017). The final forecast for very short timescales should therefore be a judicious combination, commonly known as blending, of extrapolation and NWP with data assimilation, as stated originally by Browning (1980). Several methods have been developed to blend these two sources of information (see, for example, Pierce et al., 2000, Hwang et al., 2015, Atencia et al., 2010, and Hu et al., 2019).

In broad terms, there are two ways of carrying out this blending. The first involves the introduction of NWP rainfall information into the stochastic generator. One example of this type of blending is the Short-Term Ensemble Prediction System (STEPS) (Bowler et al., 2006). STEPS is a stochastic blended precipitation system that combines an extrapolation nowcast, a downscaled NWP model and a small-scale noise field. The Spectral Prognosis (S-PROG) model developed by Seed (2003) is used in STEPS to simulate the uncertainties in the evolution of precipitation patterns. In the S-PROG model, Fourier filters are used to decompose the rainfall field into a multiplicative spectral cascade. This separation allows the different scale components of the precipitation pattern to be treated independently and attributes different weights to different scales.

The second way of blending involves assimilating the nowcast product as the external data source to establish the initial conditions of the NWP model. A good example of this type of blending was developed by Nerini et al. (2019), who used the reduced-space ensemble Kalman filter method to include the EPS forecast in the advection system.

The nowcasting mentioned in the above two types of blending is usually deterministic; however, it is also possible to use ensemble nowcasting in the blending. Blending ensemble nowcasting and ensemble NWP is a way to ensure that the transition from the nowcasting component to the NWP component is smooth and consistent in terms of the forecast. In addition, Kober et al. (2012) and Scheufele et al. (2013) have shown that, with respect to accuracy, a blending of probabilistic radar-based nowcasts and probabilistic model-based forecasts results in an overall forecast that is at least as skilful as, and may even exceed, the skill of each individual component.

A potentially important consideration in applying blending is the fact that forecast predictability varies with the meteorological situation (Stensrud and Wandishin, 2000; Kühnlein et al., 2013). Characteristics of convection such as intensity or organization are controlled in part by the large-scale flow pattern and in part by properties of the local environment such as variability in the planetary boundary layer. Done et al. (2006) introduced a convective adjustment timescale parameter to characterize the degree of large-scale versus smaller-scale/orographic forcing control by distinguishing between equilibrium and non-equilibrium convection. This parameter has been used to classify convective regimes in observational data sets (Zimmer et al., 2011) and to provide regime-dependent diagnostics of NWP models and sources of uncertainty in ensemble forecasts (Keil and Craig, 2011). In addition, this parameter can be used to modify the weighting of the nowcast (or NWP forecast) depending on the meteorological situation (Kober et al., 2014). Raynaud et al. (2015) used a scheme of blending probabilistic information by utilizing flow-dependent Bayesian weighting and time-lagged ensemble members to create a smooth transition between different initialization time runs of the NWP forecast. Raynaud's approach would be preferred over Done's parameter when time-lagged ensemble members are used.

Recent studies about the performance of different NWP forecasts for flood forecasting (Cloke et al., 2017) and about the performance of nowcasting (Berenguer et al., 2011) with respect to rainfall have shown that forecast performance varies locally. This has been translated into a new trend in the blending methodologies where localization plays a role. Three examples of this new trend are presented in Moisselin et al. (2019), in which the weights used in the blending are computed for small boxes covering the whole domain, Sideris et al. (2020), in which the growth and decay of the rainfall system is computed locally from the NWP product and introduced into the nowcast, and Atencia et al. (2020a, 2020b), in which a variational approach with localization is used to compute local and flow-dependent weights.

The advantages of taking into account localization considerations in a blended forecast are presented in Figure 15, which compares non-localized blending and localized blending to observation. The purple circle in each image clearly shows that the localized forecast correctly predicted the growth of a storm that was not predicted (and consequently not properly advected) in the non-localized forecast. Non-localized blending has a unique (global) weight for the whole domain, which forces it to find a compromise between regions where the extrapolation has correctly forecasted a storm and regions where NWP has correctly predicted the evolution of the storm. This compromise tends to smooth both the extrapolation and NWP fields (by averaging them) and reduces the high intensity in regions where either extrapolation or NWP has consistently been accurate. The introduction of localized weights allows these regions to give the proper weight (in a statistical sense) to either the extrapolation or the NWP model and to reduce the smoothing effect resulting from using a global weighting.

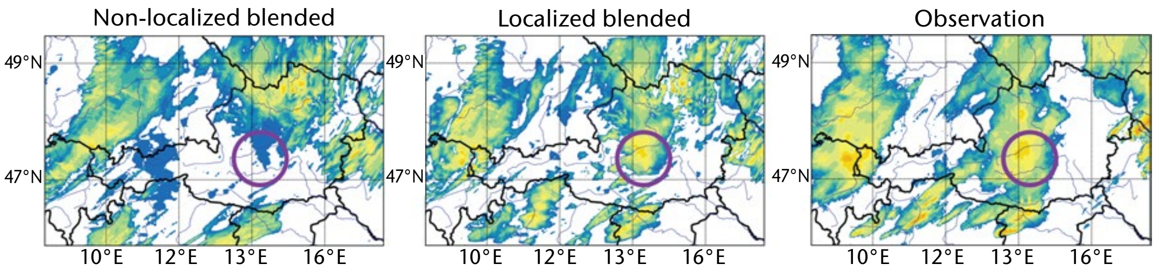

Figure 15. Comparison of classical global weight (non-localized) blending versus the new developments where the weights are local dependent (localized). The figure shows the precipitation for the two blending methods (T + 5 hours lead time) and the corresponding observation analysis (right) for 2300 UTC, 12 July 2016 to 0000 UTC, 13 July 2016. The purple circle highlights the benefits of localized blending where high intensities are present (matching the observations); the non-localized forecast, in contrast, predicts a much less significant amount of rain in this area.

Source: Aitor Atencia, Austrian Central Institute of Meteorology and Geodynamics (ZAMG)/Department of forecasting models

CHAPTER 5. MULTI-MODEL ENSEMBLES (TIER 3)

Many studies, including Vislocky and Fritsch (1995) and Whitaker et al. (2006), have found that the statistically postprocessed output from a combination of multiple prediction systems can produce a significantly more skilful forecast than the statistically postprocessed output from a single prediction system. One difficulty related to multi-model systems is that the developer must keep the training data up to date for multiple prediction systems, which can involve considerable time and effort. Further, as demonstrated in studies such as Hamill et al. (2008), Hagedorn et al. (2008), Hamill (2012), and Hamill and Scheuerer (2018), if the skill of one prediction system greatly exceeds that of other systems in a multi-model ensemble, the multi-model performance provides little improvement in the postprocessed forecast compared to the most skilful system. Nevertheless, for equally skilful systems, blending the data from multiple predictions may be an option.

Just as there are diverse approaches to postprocessing for single model systems, there are also many reliable approaches to postprocessing for multi-model systems. One such approach that has been applied in many situations involves the Bayesian model averaging (BMA) method discussed in Section 3.3.2 (Raftery et al., 2005, Sloughter et al., 2010, Wilson et al., 2007). BMA can be thought of as the regression correction of various members of the multi-model ensemble to the observations or analyses, followed by a weighted "dressing" of the regressed ensemble members. That is, the final PDF is a weighted sum of kernels of probability density centred on the regression-corrected members. As this method has been widely used, it has also been subject to considerable scrutiny, as reflected in constructive critiques such as Hamill (2007), Bishop et al. (2008), Fraley et al. (2010) and Hodyss et al. (2016). These authors suggest possible modifications to the BMA method to make it more generally extensible and to provide more realistic guidance with shorter training data sets.

Hamill et al. (2017) and Hamill and Scheuerer (2018) demonstrate another multi-model statistical method, in this case specifically for precipitation amount forecasting. As opposed to the regression-based corrections underlying BMA, in this method, biases that may depend on precipitation amount in a given model within a multi-model ensemble are compensated for using quantile mapping. Spread corrections are addressed through dressing (Hamill et al., 2017), and in a subsequent algorithmic adjustment, through a weighting of sorted members to address problems of overconfidence in the predictions.

For extremely sophisticated multi-model developers, a potential issue may be that a simple, weighted combination of postprocessed PDFs from independent prediction systems may result in an ensemble that is sub-optimal and could be sharpened, improving skill (Gneiting and Ranjan, 2013). This is worth being aware of, but it is probably a minor concern for most developers.

In summary, multi-model methods may be worth considering if developers have ready access to data from multiple ensemble prediction systems, if they are willing to maintain the multiple forecast training data sets, and if one system within the multi-model ensemble is not clearly better than the others. It is recommended that developers consult with one or multiple designated Global Data-processing and Forecasting System (GDPFS) centres which conduct ensemble forecasts for guidance on conducting multi-model ensemble postprocessing. These centres are listed in the *Manual on the Global Data-processing and Forecasting System* (WMO-No. 485) Part III.

CHAPTER 6. VERIFICATION AND VALIDATION

6.1 Validation best practice

It is important to assess the quality of any forecasting system, and an evaluation of postprocessed forecasts will demonstrate the benefits of the postprocessing method used. This is valuable information for users and essential for monitoring the ongoing quality of the product.

The NWP ensemble forecast data provided to WMO members via the WMO Information System (WIS) is quality-controlled on a continuous basis by the producing centres. A standard set of verification scores is used in the quality control process, and the results are available on the websites of the WMO Lead Centre for Coordination of Deterministic NWP Verification and the Lead Centre for Coordination of EPS Verification:

https://apps.ecmwf.int/wmolcdnv/

http://epsv.kishou.go.jp/EPSv/.

It is recommended that the same verification scores should be used to evaluate postprocessed forecasts. Details of the scores and how to compute them are given on the Lead Centre websites.

More information about a wide range of methods to evaluate forecasts, as well as information about the verification scores used by the Lead Centres, is provided by the WMO Joint Working Group on Forecast Verification Research on the following website:

https://www.cawcr.gov.au/projects/verification/.

The following books also contain comprehensive information about forecast verification:

Jolliffe, I.T. and D.B. Stephenson, 2012: *Forecast Verification: A Practitioner's Guide in Atmospheric Science.* 2nd Edition. Chichester: Wiley. https://onlinelibrary.wiley.com/doi/book/10.1002/9781119960003.

Wilks, D.S., 2019*: Statistical Methods in the Atmospheric Sciences*. Fourth edition. Amsterdam, Elsevier. ISBN 9780128158234, eBook ISBN 9780128165270.

6.2 Metrics for deterministic forecasts

For deterministic postprocessing such as for the ensemble mean, it is recommended that the standard verification measures such as the root mean square error (RMSE) should be used. Details of the RMSE and other recommended scores are provided on the website of the Lead Centre for Coordination of Deterministic NWP Verification:

https://confluence.ecmwf.int/display/WLD/Score+definitions+and+requirements.

6.3 Metrics for probabilistic forecasts

6.3.1 *Introduction*

Ensemble forecast systems were first designed in the 1990s in order to provide guidance for numerical probabilistic forecasts. Subsequently, metrics to evaluate ensemble-based probabilistic forecasts were developed to measure the forecast performance for general users, decision makers and model developers. The verification system mainly focuses on two attributes of the forecast: reliability and resolution (Jolliffe and Stephenson, 2012; Toth et al., 2006); this is in contrast to the traditional verification measures for deterministic and ensemble mean forecasts, which include pattern anomaly correlation (PAC), RMSE and absolute error (ABS).

A probabilistic forecast gives a probability of an event or category occurring, with a value between 0 and 1 (or 0% and 100%). Usually, it is difficult to verify a single probabilistic forecast. Instead, a set of probabilistic forecasts, p(i), is verified using observations (or best analyses), confirming that those events either occurred (o(i)=1) or did not occur (o(i)=0). Practically, an event or category could be defined based on 1) user-defined thresholds, 2) climatological percentile (or category) or climatological-equal-likely percentiles/bins, or 3) the ranking of the ensemble members (with an equal weighting of the members).

With respect to the methodology of probabilistic verifications, "reliability" is defined as the agreement between forecast probability and mean observed frequency; "resolution" is defined as the ability of the forecast to resolve the set of sample events into subsets with characteristically different outcomes; "sharpness" is defined as the tendency to forecast probabilities near 0 or 1, as opposed to values clustered around the mean; and "uncertainty" is defined as the nature (in other words, climatological frequency) of a specified event or category.

6.3.2 *Methodology of verification*

1) Brier score:

The Brier score (BS) (see Brier, 1950 and Wilks, 2019) can be defined as:

$$BS = \frac{1}{n}\sum_{k=1}^{n}(p_k - o_k)^2 \qquad 6.1$$

where p is the forecast probability, o is the observed frequency, and the index k denotes the number of the n forecast events/pairs. p and o are limited from 0 to 1 in the probability sense. $BS = 0$ indicates a perfect forecast, and $BS = 1$ indicates a forecast that does not predict any of the observed events.

Through a common algebraic decomposition, BS can be expressed as three separate terms:

$$BS = \frac{1}{n}\sum_{i=1}^{I} N_i(p_i - \overline{o_i})^2 - \frac{1}{n}\sum_{i=1}^{I} N_i(\overline{o_i} - \overline{o})^2 + \overline{o}(1 - \overline{o}) \qquad 6.2$$

where n is the total forecast number issued, I is the number of unique forecasts issued and N_i is the number of forecasts with the same probability category. The conditional probability of the observed and sample climatology can be expressed as follows:

$$\overline{o_i} = p(o_i \mid p_i) = \frac{1}{N_i}\sum_{k \in N_i} o_k \text{ and } \overline{o} = \frac{1}{n}\sum_{k=n}^{n} o_k \qquad 6.2.1$$

Equation 6.2 can be summarized as BS = reliability + resolution + uncertainty. Statistical postprocessing typically corrects reliability. Resolution is generally harder to improve through postprocessing and indeed, sometimes can be adversely affected by postprocessing.

In addition, when considering forecast application through calibration, BS can be expressed as two terms: calibration + refinement:

$$BS = \frac{1}{n}\sum_{i=1}^{I} N_i(y_k - \overline{o_i})^2 - \frac{1}{n}\sum_{i=1}^{I} N_i(\overline{o_i}(1 - \overline{o_i})) \qquad 6.3$$

Calibration (the first term) can be used as a measure of statistical calibration and is equal to reliability. Refinement (the second term) is an aggregation of resolution and uncertainty and is related to the area under the relative operating characteristics (ROC) curve.

When a good reference BS is used, the Brier skill score (BSS) may be generated from the following formula:

$$BSS = \frac{BS_f - BS_{ref}}{BS_{perf} - BBS_{ref}} = 1 - \frac{BS_f}{BS_{ref}} \qquad 6.4$$

where *ref* is the reference, which is mostly the climatology (the forecast probability equals the climatological frequency of the event), but may be user-specified, and $BS_{perf} = 0$ indicates a perfect forecast. BSS can be simplified as the right side of equation 6.4, which ranges from 0 to 1.

Figure 16 demonstrates the reliability diagram with respect to the Brier score and its decomposition. The x-axis indicates the forecast probability, and the y-axis indicates the observed relative frequency. The diagonal solid straight line represents perfect reliability. The reliability of the forecast is determined by how closely the blue curve deviates from the diagonal line. If the blue curve is below the diagonal line, the forecast probabilities are too high (over-forecast); if the blue curve is above the diagonal line, the forecast probabilities are too low (under-forecast). As the curve flattens, the resolution decreases. The histogram shows the forecast frequency for each probability bin and the sharpness of the forecast.

2) Ranked probability score:

The ranked (ordered) probability score (RPS) verifies multi-category probability forecasts in order to measure both reliability and resolution based on defined (selected) climatologically equally likely bins:

$$RPS = 1 - \frac{1}{k-1}\left[\sum_{i=1}^{k}\left(\sum_{n=1}^{i} P_n - \sum_{n=1}^{i} O_n\right)^2\right] \qquad 6.5$$

where k (note that k is not equal to 1) is the total category which has been defined and n is a number from 1 to k. P and O represent forecast and observation (or best analysis) probability, respectively. When a reference RPS (RPS_{ref}) (usually climatology) is defined, a ranked probability skill score (RPSS) can be converted from the following formula:

$$RPSS = \frac{RPS_f - RPS_{ref}}{RPS_{perf} - RPS_{ref}} = 1 - \frac{RPS_f}{RPS_{ref}} \qquad 6.6$$

perf indicates a perfect forecast, and RPS_{perf} is usually defined as 1. $RPSS$ can be simplified as the right side of the equation and ranges from 0 to 1.

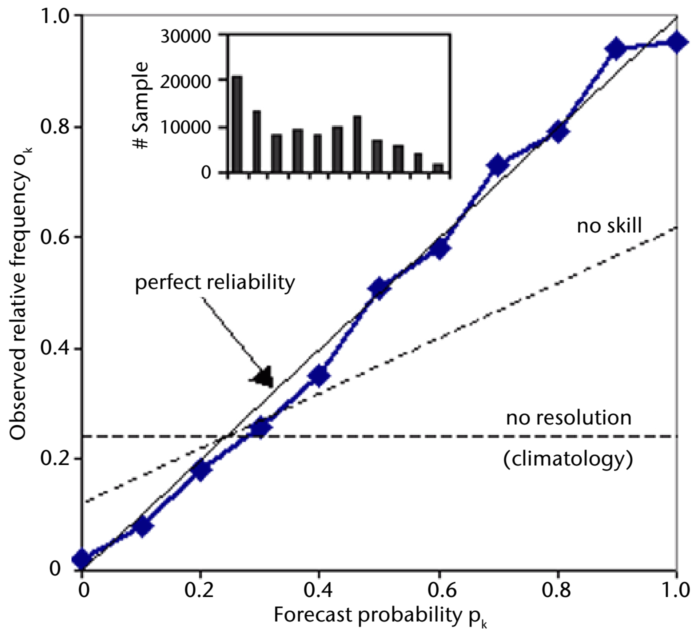

Figure 16. Reliability diagram demonstrating the relationship between forecast probability and observed relative frequency. When the no resolution and no skill (with respect to the climatology) lines are included, this diagram is also called an attributes diagram.

Source: https://www.cawcr.gov.au/projects/verification/

6.3.3 *Continuous ranked probability score*

The continuous ranked probability score (CRPS) is a much-used measure of performance for probabilistic forecasts of a scalar observation. It is a quadratic measure of the difference between the forecast cumulative distribution function (CDF) and the empirical CDF of the observation. It is similar to the RMSE measure but is used for the forecast accumulated distribution. *CRPS* can be represented in the following manner:

$$CRPS = \int_{-\infty}^{+\infty} \left[F(x) - H(x - x_0)\right]^2 dx \qquad 6.7$$

where $F(x)$ represents the forecast distribution and $H(x)$ is the Heaviside function, further described below:

$$H\left(x - x_0\right) = \begin{cases} 0, x \le x_0 \\ 1, x > x_0 \end{cases} \qquad 6.7.1$$

A calculation of the forecast CRPS is presented in Figure 17. To produce this figure, 10 ensemble members were ranked from low to high according to value, with equal weighting, on the x-axis. The actual observation is marked in blue. The percentage (0%,10%, 20%…100%) is marked on the y-axis.

In the same way that BS can be decomposed (see Section 6.3.2), CRPS can also be decomposed into reliability, resolution and uncertainty. The continuous ranked probability skill score (CRPSS) can be expressed as:

$$CRPSS = 1 - \frac{CRPS_f}{CRPS_{ref}} \qquad 6.8$$

where $CRPS_{ref}$ is a CRPS reference that is usually calculated using climatological data. *CRPSS* ranges from 0 to 1.

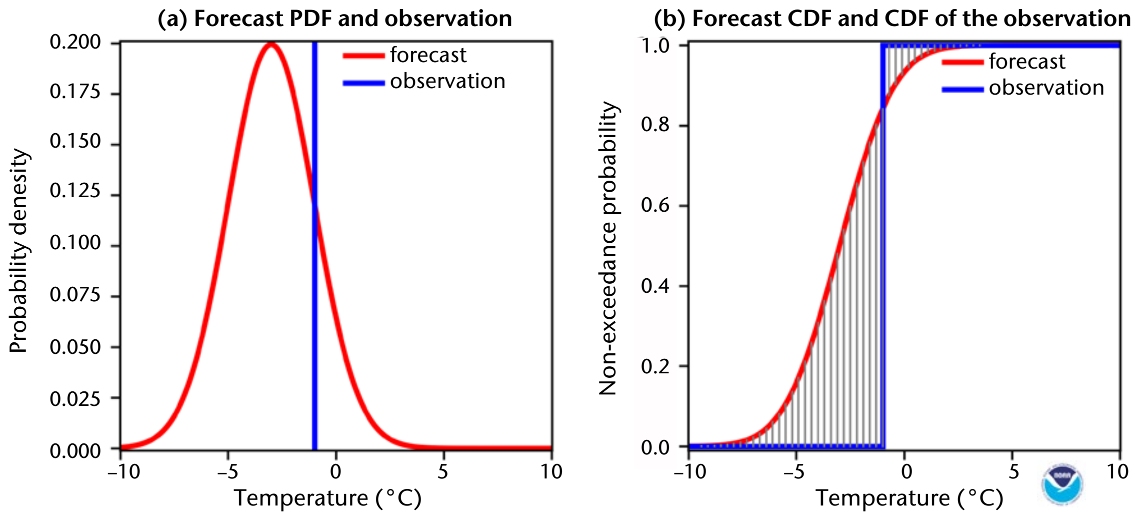

Figure 17. Schematic diagram demonstrating an ensemble forecast distribution for temperature. Left: forecast PDF and observation value; right: forecast CDF and CDF of the observation. The CRPS calculation is an integration of the shaded area bounded by the forecast CDF and the CDF of the observation in the right image.

Source: T. Hamill, NOAA/Office of Oceanic and Atmospheric Research/Physical Sciences Laboratory

CHAPTER 7. DATA INFORMATION

7.1 Data-science considerations: the bias-variance trade-off

Fundamental to an understanding of all statistical algorithms is the "bias-variance trade-off". There is an extensive discussion of this in Hamill (2018), and hence a long discussion here is omitted. With a given amount of training data, more complicated statistical algorithms generally are capable of reducing bias with respect to observations, but this comes at the expense of an increase in error variance (error of the forecasts when validated after the fact). Conversely, it is possible to optimize for reduced error, but this optimization results in larger biases. Model development often amounts to choices as to how much bias versus how much variance the developer is willing to accept given the training data at hand. While trade-offs between bias and variance are somewhat inevitable, generally more training data (a longer time series of past forecasts and observations) will allow more sophisticated techniques capable of limiting both bias and variance to be developed.

In order to increase skill, it may be necessary for some forecast variables to have a longer training data time series than others. A consistent short-term forecast temperature bias that does not vary much from one day to the next usually can be corrected adequately with relatively short training data sets. If biases vary widely across training samples, or if the biases are highly state-dependent (for example, if the biases are different for high precipitation amounts than for low precipitation amounts), then more training samples are needed to produce guidance of acceptable quality. If the forecasts are for unusual features such as very heavy precipitation, or if they are for time averages, such as precipitation from three to four weeks lead time, then often retrospective forecasts spanning multiple years or even decades are needed to provide sufficient samples (Hamill et al., 2004; Scheuerer and Hamill, 2018). For example, if an 8- to 14-day lead time-averaged probabilistic temperature forecast uses forecasts initialized on the previous 60 days for training data, a forecast initialized on 1 December (valid from 9 to 15 December) will have a significant amount of data that overlaps with data generated from a forecast initialized on 2 December (valid from 10 to 16 December). In other words, if only data from the previous 60 days are used, there then are only 60/7 reasonably independent samples, clearly not enough to reliably estimate the systematic error among the large random errors that exist for 8- to 14-day forecasts.

Training data can be increased by using reforecasts, or in some circumstances, by using methods that pool the training data in defensible ways. For example, if two locations A and B are suspected of having similar bias characteristics but are far enough apart to provide quasi-independent samples, the data from these two locations may be pooled to increase the training sample size. For a more detailed explanation of this point, see the discussion of "supplemental locations" in Hamill et al. (2017).

7.2 Data management considerations

If statistical algorithms are applied to improve forecasts, the developer must obtain and maintain archives of past forecasts and observations/gridded analyses.

Determining the length and variety of retrospective forecast data that a developer should apply depends on many factors. One of these concerns the data, storage, and data transfer capabilities that are available, as well as the time that is required for their management. Generally, if these capabilities are limited, the developer should choose a methodology that uses shorter, simpler training data sets and simpler algorithms, with the understanding that the benefit provided will not be as significant as the benefit that would be obtained from algorithms using larger data sets and more complicated methodologies.

If the developer chooses not to regularly update the forecast training data, the quality of the resulting statistically postprocessed guidance may be degraded. For example, if the training data has been archived from an older forecast model version that has a cold bias and the prediction centre has recently upgraded the model to one with a slight warm bias, using the older forecasts to diagnose the statistical corrections and then applying those corrections to the new, real-time forecasts will actually degrade the postprocessed product quality.

In addition, the developer of the method may be interested in applying multi-model techniques. This requires past and current forecast guidance from multiple prediction centres. Each prediction centre commonly applies its own model independently of the others. It is thus desirable for the developer to carefully track the model upgrades and to update the training data relevant for each constituent prediction. Clearly, there is potential for major confusion in keeping the training data up to date for each prediction system so that it is possible to produce high-quality guidance.

With respect to refreshing the forecast training data, if, for example, the developer is using a reforecast data set (retrospective forecasts of the weather using the current operational model version, see Hamill et al., 2013 and Vitart et al., 2019), the most recent version of the reforecast should be used with the upgrade to a new model. In some centres, such as NCEP in the US, these reforecasts are pre-computed and made available prior to the change of model version. Other centres, such as ECMWF, have a rolling-window production strategy; for example, in February they may compute retrospective forecasts for dates in March over the previous 20 years. In this way, many years of January–February–March data are available to develop a statistical model appropriate in the current February.

Because of the challenges in maintaining up-to-date reforecast data, many developers may choose to only archive operational forecasts as they become available. In this scenario, statistical postprocessing is only applied to the last few months of forecasts for training data, and the user accepts the reduced quality of the postprocessed product until such time as the last few months of training data consist of data from the new model version.

Some current data management challenges may, in the near future, become less challenging if computations are done in the “cloud”, that is, on major computer clusters run by large organizations such as Amazon, Microsoft or IBM. Several NWP providers as of 2020 had migrated or were in the process of migrating their forecast data from local servers to the cloud. For developers choosing cloud-based development, data management may become simpler, and the relevant weather services may refresh the forecast data sets themselves. The downsides of this approach are that the different centres may use different cloud providers and the developer will be paying the cloud service provider for use of their computers and for downloading the data to the developer’s own local storage. Still, this may be less expensive in the long run than paying the salary of a worker to maintain and regularly update the local storage of recent forecast data.

Another consideration is whether developers have a long or short time series of local, quality-controlled observations that they will use to statistically adjust the forecasts, or whether developers are completely lacking these time series. A long time series of forecasts is not very useful if the developer does not have a correspondingly long time series of observations or analyses. Observations should, wherever possible, be quality-controlled before use as a small number of erroneous observations could greatly reduce the quality of statistical corrections or even lead to the degradation of the raw forecasts. If developers have no observations or gridded analyses available, they may choose to use methods to synthesize information from recent forecasts (see Sections 7.3.1 and 7.3.2 below) or to access observations or analyses produced by others. For example, the European Copernicus service provides global reanalyses of many variables of interest (temperature, wind, precipitation) that, while not perfect in their quality, may still be useful in the statistical adjustment of forecasts.

Some of the algorithms discussed in this document may use data other than forecasts and observations. For example, if a developer is making forecasts for a region with many mountains, valleys and lakes, ancillary data sets might include terrain elevation data, land/water masks, or land-surface characteristics (urban, farmland, forests, and so forth). These data will aid in the development of algorithms that downscale the coarser resolution forecast data to a higher-resolution analysis grid.

7.3 Basic data characteristics of ensemble forecasts

7.3.1 *Lagged ensembles*

For a particular variable, such as total precipitation amount, a user may have a number of forecasts which are valid at the same time. For example, today's 12- to 24-hour forecast guidance for precipitation forecasts the precipitation for the same time period as yesterday's 36- to 48-hour forecast guidance for precipitation. Yesterday's forecast may be less accurate, but in practice, it has been shown that a weighted, lagged-average forecast, even without statistical postprocessing, commonly produces a forecast with a lower RMSE than the most recent forecast alone (Hoffman and Kalnay, 1983). Postprocessing to generate a weighted, lagged-average forecast is the simplest type of postprocessing that can be done to provide some skill improvement, and it may be applied in the absence of observational data. As is common with any weighted combination, though the resulting forecast may reduce the mean error, the weighted-mean guidance has the tendency to reduce the amplitude of the maxima and enlarge the area with non-zero precipitation (see Figure 18) – another example of the bias-variance trade-off. This is because the peak precipitation is often in slightly different locations in each lagged forecast.

7.3.2 *Multi-model ensemble combinations*

In recent years, attention has turned to how to optimally combine predictions from several operational global ensemble prediction systems. Can a benefit be obtained from generating a super-ensemble consisting of the union of the members from all the constituent prediction systems? Yes, but it is important to understand what may be sub-optimal about such a union. Commonly, these ensemble systems exhibit different skills and biases. For example, perhaps, for a chosen region, ECMWF's forecasts are commonly too warm on average, while the UK Met Office's forecasts are commonly too cold. Perhaps ECMWF's forecasts exhibit relatively the flat rank histograms (Hamill, 2001) of a system that makes rather reliable probability forecasts, while NCEP's rank histograms exhibit the characteristic U-shape of an overconfident, unreliable prediction system. An examination of a combination of members from each system's raw ensemble may reveal evidence of clustering by system, perhaps with peaks in ensemble

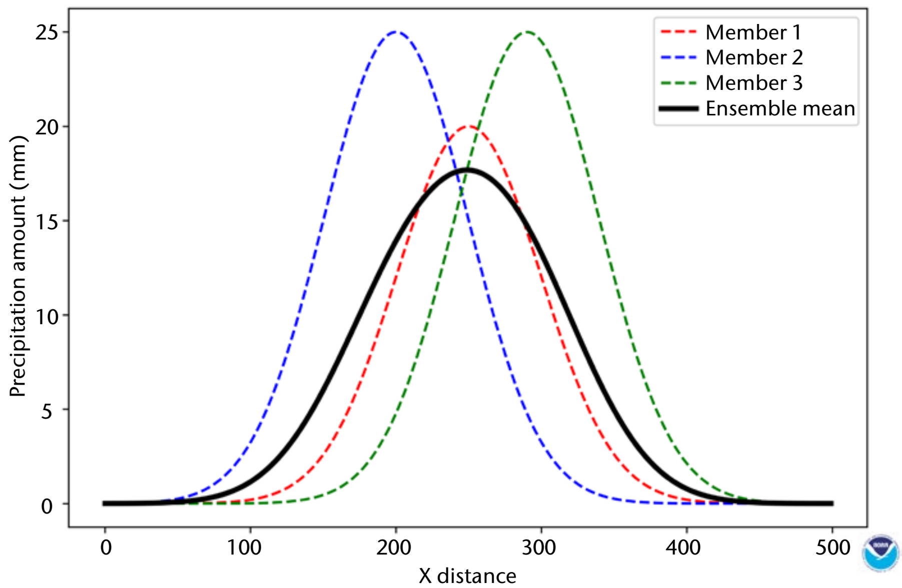

Figure 18. A synthetic example of a three-member ensemble of precipitation amounts (coloured dashed lines) and the ensemble mean (thick black line) along a latitude circle

Source: T. Hamill, NOAA/Office of Oceanic and Atmospheric Research/Physical Sciences Laboratory relative frequency around ECMWF, UK Met Office, and US NWS systems. This is not a reflection of the true multi-modality of the forecast uncertainty, but rather of the prediction-system dependent biases.

If one of each model in a multi-model system is considered, then a combination of postprocessed guidances from multiple models is typically more skilful (Vislocky and Fritsch, 1995; Whitaker et al., 2006) than a single model alone. An exception is when one model is clearly more skilful than the others (Hamill, 2012); in such a circumstance, the most skilful model alone may be sufficient.

In summary, paying attention to the maintenance of the data used in the method is as important as the selection of the algorithm. In order to produce operational postprocessed guidance, developers will need to regularly update to new training data so that they can train new statistical models using that new data. This will take some significant time to do properly, and developers should plan for this.

7.4 **Use of and need for reforecasts for calibration**

Reforecasts have been generated for several operational models, with descriptions in Hamill et al. (2004), Hagedorn (2008) Gagnon et al. (2013), and Hamill et al. (2013). These are retrospective numerical forecasts of the weather constructed to the extent possible to be consistent with the operational forecast system. Reforecasts can be particularly helpful in the statistical postprocessing of unusual events that may not occur in a short training sample, and they are also helpful for improving the skill of postprocessed events that represent time averages, such as the average over the second week of the forecast.

The reforecast is also used to compute the model climate for weeks 1 to 4 and the monthly forecast of temperature and precipitation anomalies (see Lin et al., 2016). The reforecast is further used to compute the model climate distribution in order to calculate the extreme forecast index (see Lin, 2012).

CHAPTER 8. SOFTWARE AND TECHNICAL INFORMATION

8.1 Data sources

- WMO-designated GDPFS centres (World Meteorological Centres (WMCs) and Regional Specialized Meteorological Centres (RSMCs)) make available deterministic and ensemble forecast data via WIS. Some observations are also available through WIS.

- The THORPEX Interactive Grand Global Ensemble (TIGGE) archive (https://confluence.ecmwf.int/display/TIGGE) provides access to a large archive of data from global EPSs which may be used for testing methods and training systems.

- Several global NWP centres, including NCEP (USA), make EPS forecasts freely available via the Internet. It is likely that in the future, more centres will make both real-time and archive data available on cloud servers.

8.2 Computing platforms

Many of the methods described as **Tier 1** can be implemented on relatively small IT systems using relatively small quantities of data. **Tier 2** and **Tier 3** methods generally require access to larger capacity systems with the ability to handle large quantities of data.

Increasingly, both NWP and observational data are becoming available on public cloud computing systems. This provides a good opportunity for NMHSs which may not have access to high-performance computing facilities to do EPS postprocessing in the cloud without having to download large quantities of data and therefore without the need for high bandwidth connections.

8.3 R packages

R is an open-source, freely available statistical software language and environment which supports programming tools for a variety of applications. These include statistical techniques for ensemble forecasts, such as the previously discussed EMOS and MBMP. In addition, R also provides tools for data processing, graphical representation and verification. R is available from the Comprehensive R Archive Network (CRAN). Jakob W. Messner's excellent chapter "Ensemble Postprocessing with R" in the book *Statistical Postprocessing of Ensemble Forecasts* (see the References section for full bibliographic details), describes the use of toolboxes available in R for statistical postprocessing. This chapter reviews the available packages for statistical postprocessing, including verification packages, and provides examples of the figures representing the corresponding output. The different packages available as of 2018 are listed in Appendix B of the chapter. Frequently used packages include ensembleMOS (Yuen et al., 2017) and ensembleBMA (Fraley et al., 2018), while packages used for verification include SpecsVerification (Siegert, 2017) and scoringRules (Jordan et al., 2017).

More recent packages have also been developed and can be found at https://cran.r-project.org/web/packages/available_packages_by_date.html. Recent packages include CSTools (Perez-Zanon et al., 2019).

8.4 Python libraries

There are many Python libraries which can be used in the postprocessing of NWP data. Examples of libraries specifically designed for use with meteorological data include:

IMPROVER: This is an open-source library of postprocessing code developed by the UK Met Office in collaboration with the Australian Bureau of Meteorology. It is designed to work in a probability framework and includes routines specifically designed to exploit convective scale models and ensembles and for blending multiple models and ensembles: https://github.com/metoppv/improver.

Community Atmospheric Model Postprocessing System (**CAMPS**): This is a software infrastructure that supports the statistical postprocessing (StatPP) of atmospheric data and is maintained as community code. CAMPS is currently under development and managed by the Meteorological Development Lab (MDL) of NOAA in the USA. It includes standards and tools for data representation as well as software repositories that facilitate the use of these standards.

CAMPS aims to both modernize MDL's StatPP infrastructure and make it readily accessible to outside users. The long-term vision for CAMPS is to fully replace MDL's MOS and Localized Aviation MOS Program, support the National Blend of Models, and allow for streamlined additions of the most cutting-edge StatPP techniques.

The latest version of CAMPS can be found on GitHub (https://github.com/NOAA-MDL/CAMPS). Version 1.0.0 of CAMPS was released in January 2020. Although version 1.0.0 has limited capabilities, new releases with expanded capabilities are forthcoming. It is recommended that users check the GitHub page frequently for the latest updates and version of the software.

More general code libraries for analysing and visualising Earth science data include Iris and Xarray:

Iris is an open-source library supporting data in CF-NetCDF and other formats, including GRIB2, led by the UK Met Office: https://scitools.org.uk/iris/docs/latest and https://github.com/SciTools/iris.

Xarray is a library that is more flexible and higher performance than Iris, and compared to Iris, enforces the CF convention in NetCDF less tightly. It is open source (Apache 2 licence) and publicly available. For more information, see http://xarray.pydata.org/en/stable/ or the repository at https://github.com/pydata/xarray.

8.5 GrADS

The Grid Analysis and Display System (GrADS) is an interactive desktop tool that is used for easy access, manipulation and visualization of Earth science data. GrADS has two data models for handling gridded and station data. It supports many data file formats, including binary (stream or sequential), GRIB (versions 1 and 2), NetCDF, HDF (versions 4 and 5), and BUFR (for station data). GrADS has been implemented worldwide on a variety of commonly used operating systems and is freely distributed over the Internet: http://cola.gmu.edu/grads/.

LIST OF ACRONYMS

BMA	Bayesian Model Averaging
CDF	Cumulative Distribution Function
ECC	Ensemble Copula Coupling
ECMWF	European Centre for Medium-Range Weather Forecasts
EcPoint	ECMWF point rainfall probability
EFI	Extreme Forecast Index
EMOS	Ensemble Model Output Statistics
EVMOS	Error-in-Variables Model Output Statistics
EPS	Ensemble Prediction System
GDPFS	Global Data-processing and Forecasting System
MBMP	Member-by-member postprocessing
MOS	Model Output Statistics
NCEP	National Centers for Environmental Prediction
NGR	Non-homogeneous Gaussian Regression
NHR	Non-homogeneous Regression
NMHS	National Meteorological and Hydrological Service
NWP	Numerical Weather Prediction
PDF	Probability Density Function
QRF	Quantile Regression Forest
RF	Random Forest
RMSE	Root Mean Square Error
RSMC	Regional Specialized Meteorological Centre
SAMOS	Standardized Anomaly Model Output Statistics
WIS	WMO Information System
WMC	World Meteorological Centre

REFERENCES

Andersson, T. and K.-I. Ivarsson. 1991. “A Model for Probability Nowcasts of Accumulated Precipitation Using Radar”. *Journal of Applied Meteorology and Climatology* 30(1). American Meteorological Society: 135–141. doi: 10.1175/1520-0450(1991)030<0135:AMFPNO>2.0.CO;2.

Atencia, A., T. Rigo, A. Sairouni, J. Moré, J. Bech, E. Vilaclara, J. Cunillera, M.C. Llasat and L. Garrote. 2010. “Improving QPF by Blending Techniques at the Meteorological Service of Catalonia”. *Natural Hazards and Earth System Sciences* 10(7). Copernicus GmbH: 1443–1455. doi: 10.5194/nhess-10-1443-2010.

Atencia, A., Y. Wang, A. Kann and F. Meier. 2020a. “Localization and Flow-Dependency on Blending Techniques”. *Meteorologische Zeitschrift*. Schweizerbart’sche Verlagsbuchhandlung: 231–246. doi: 10.1127/metz/2019/0987.

Atencia, A., A. Kann, Y. Wang and F. Meier. 2020b. “Localized Variational Blending for Nowcasting Purposes”. *Meteorologische Zeitschrift* 29: 247–261. doi: 10.1127/metz/2020/1003.

Baran, S. 2014. “Probabilistic Wind Speed Forecasting Using Bayesian Model Averaging with Truncated Normal Components”. *Computational Statistics & Data Analysis* 75: 227–238. doi: 10.1016/j.csda.2014.02.013.

Baran, S. and S. Lerch. 2018. “Combining Predictive Distributions for the Statistical Post-Processing of Ensemble Forecasts”. *International Journal of Forecasting* 34(3). Elsevier: 477–496. doi: https://doi.org/10.1016/j.ijforecast.2018.01.005.

Berenguer Ferrer, M., D. Sempere Torres and G. Pegram. 2011. “SBMcast - An Ensemble Nowcasting Technique to Assess the Uncertainty in Rainfall Forecasts by Lagrangian Extrapolation”. *Journal of Hydrology* 404(3–4). Elsevier Science Direct: 226–240. doi: 10.1016/j.jhydrol.2011.04.033.

Bishop, C.H. and K.T. Shanley. 2008. “Bayesian Model Averaging’s Problematic Treatment of Extreme Weather and a Paradigm Shift That Fixes It”. *Monthly Weather Review* 136(12). American Meteorological Society: 4641–4652. doi: 10.1175/2008MWR2565.1.

Bourgouin, P. 2000. “A Method to Determine Precipitation Types”. *Weather and Forecasting* 15(5). American Meteorological Society: 583–592. doi: 10.1175/1520-0434(2000)015<0583:AMTDPT>2.0.CO;2.

Bowler, N.E., C.E. Pierce and A.W. Seed. 2006. “STEPS: A Probabilistic Precipitation Forecasting Scheme Which Merges an Extrapolation Nowcast with Downscaled NWP”. *Quarterly Journal of the Royal Meteorological Society* 132(620): 2127–2155. doi: https://doi.org/10.1256/qj.04.100.

Bremnes, J.B. 2004. “Probabilistic Forecasts of Precipitation in Terms of Quantiles Using NWP Model Output”. *Monthly Weather Review* 132(1). American Meteorological Society: 338–347. doi: 10.1175/1520-0493(2004)132<0338:PFOPIT>2.0.CO;2.

Brier, G.W. 1950. “Verification Of Forecasts Expressed In Terms Of Probability”. *Monthly Weather Review* 78(1). American Meteorological Society: 1–3. doi: 10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2.

Browning, K.A. 1980. “Review Lecture: Local Weather Forecasting”. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences* 371(1745). The Royal Society: 179–211. doi: https://doi.org/10.1098/rspa.1980.0076.

Casaioli, M., R. Mantovani, F. Proietti Scorzoni, S. Puca, A. Speranza and B. Tirozzi. 2003. “Linear and Nonlinear Post-Processing of Numerically Forecasted Surface Temperature”. *Nonlinear Processes in Geophysics* 10(4/5). Copernicus GmbH: 373–383. doi: 10.5194/npg-10-373-2003.

Cloke, H.L., F. Pappenberger, P.J. Smith and F. Wetterhall. 2017. “How Do I Know If I’ve Improved My Continental Scale Flood Early Warning System?” *Environmental Research Letters* 12(4). IOP Publishing: 044006. doi: 10.1088/1748-9326/aa625a.

Cui, B., Z. Toth, Y. Zhu and D. Hou. 2012. “Bias Correction for Global Ensemble Forecast”. *Weather and Forecasting* 27(2). American Meteorological Society: 396–410. doi: 10.1175/WAF-D-11-00011.1.

Dabernig, M., G.J. Mayr, J.W. Messner and A. Zeileis. 2017. “Spatial Ensemble Post-Processing with Standardized Anomalies”. *Quarterly Journal of the Royal Meteorological Society* 143(703): 909–916. doi: https://doi.org/10.1002/qj.2975.

Dabernig, M., I. Schicker, A. Kann, Y. Wang and M.N. Lang. 2020. “Statistical Post-Processing with Standardized Anomalies Based on a 1 Km Gridded Analysis”. *Meteorologische Zeitschrift*. Schweizerbart’sche Verlagsbuchhandlung: 265–275. doi: 10.1127/metz/2020/1022.

Delle Monache, L., T. Nipen, Y. Liu, G. Roux and R. Stull. 2011. “Kalman Filter and Analog Schemes to Postprocess Numerical Weather Predictions”. *Monthly Weather Review* 139: 3554–3570. doi: 10.1175/2011MWR3653.1.

Delle Monache, L., F.A. Eckel, D.L. Rife, B. Nagarajan and K. Searight. 2013. “Probabilistic Weather Prediction with an Analog Ensemble”. *Monthly Weather Review* 141(10). American Meteorological Society: 3498–3516. doi: 10.1175/MWR-D-12-00281.1.

Djuric, D. 1994. *Weather Analysis*. 1st edition. Englewood Cliffs, New Jersey. Prentice Hall.

Done, J.M., G.C. Craig, S.L. Gray, P.A. Clark and M.E.B. Gray. 2006. "Mesoscale Simulations of Organized Convection: Importance of Convective Equilibrium". *Quarterly Journal of the Royal Meteorological Society* 132: 737–756. doi: 10.1256/qj.04.84.

European Centre for Medium-Range Weather Forecasts: Forecast User Guide. https://confluence.ecmwf.int/display/FUG/Forecast+User+Guide.

Fraley, C., A.E. Raftery and T. Gneiting. 2010. "Calibrating Multimodel Forecast Ensembles with Exchangeable and Missing Members Using Bayesian Model Averaging". *Monthly Weather Review* 138(1). American Meteorological Society: 190–202. doi: 10.1175/2009MWR3046.1.

Fraley, C., A. E. Raftery, J. M. Sloughter and T. Gneiting. 2018. "Probabilistic Forecasting using Ensembles and Bayesian Model Averaging". https://cran.r-project.org/web/packages/ensembleBMA/ensembleBMA.pdf.

Gagnon, N., H. Lin, S. Beauregard, M. Charron, B. Archambault, R. Lahlou and C. Côté. 2013. "Improvements to the Global Ensemble Prediction System (GEPS) from version 3.0.0 to version 3.1.0". Development and Operations Divisions. Meteorological Research Division at the Canadian Meteorological Center of Environment Canada. http://collaboration.cmc.ec.gc.ca/cmc/CMOI/product_guide/docs/lib/technote_geps310_20131204_e.pdf.

Galway, J.G. 1956. "The Lifted Index as a Predictor of Latent Instability". *Bulletin of the American Meteorological Society* 37(10). American Meteorological Society: 528–529. doi: 10.1175/1520-0477-37.10.528.

George J.J. 1960. Weather Forecasting for Aeronautics. New York. Academic Press. https://www.elsevier.com/books/weather-forecasting-for-aeronautics/george/978-1-4832-3320-8.

Germann, U. and I. Zawadzki. 2004. "Scale Dependence of the Predictability of Precipitation from Continental Radar Images. Part II: Probability Forecasts". *Journal of Applied Meteorology and Climatology* 43(1). American Meteorological Society: 74–89. doi: 10.1175/1520-0450(2004)043<0074:SDOTPO>2.0.CO;2.

Germann, U., I. Zawadzki and B. Turner. 2006. "Predictability of Precipitation from Continental Radar Images. Part IV: Limits to Prediction". *Journal of the Atmospheric Sciences* 63(8). American Meteorological Society: 2092–2108. doi: 10.1175/JAS3735.1.

Glahn, B. 2014. "A Nonsymmetric Logit Model and Grouped Predictand Category Development". *Monthly Weather Review* 142: 2991–3002. doi: 10.1175/MWR-D-13-00300.1.

Glahn, B., M. Peroutka, J. Wiedenfeld, J. Wagner, G. Zylstra, B. Schuknecht and B. Jackson. 2009. "MOS Uncertainty Estimates in an Ensemble Framework". *Monthly Weather Review* 137(1). American Meteorological Society: 246–268. doi: 10.1175/2008MWR2569.1.

Glahn, H.R. and D.A. Lowry. 1972. "The Use of Model Output Statistics (MOS) in Objective Weather Forecasting". *Journal of Applied Meteorology* 11: 1203–1211. doi: 10.1175/1520-0450(1972)011<1203:TUOMOS>2.0.CO;2.

Gneiting, T. and R. Ranjan. 2013. "Combining Predictive Distributions". *Electronic Journal of Statistics* 7(none). Institute of Mathematical Statistics and Bernoulli Society: 1747–1782. doi: 10.1214/13-EJS823.

Gneiting, T., A.E. Raftery, A.H. Westveld and T. Goldman. 2005. "Calibrated Probabilistic Forecasting Using Ensemble Model Output Statistics and Minimum CRPS Estimation". *Monthly Weather Review* 133(5). American Meteorological Society: 1098–1118. doi: 10.1175/MWR2904.1.

Golding, B.W. 1998. "Nimrod: A System for Generating Automated Very Short Range Forecasts". *Meteorological Applications* 5(1): 1–16. doi: https://doi.org/10.1017/S1350482798000577.

Goodfellow, I., Y. Bengio and A. Courville. 2016. *Deep Learning*. Illustrated edition. Cambridge, Massachusetts. The MIT Press.

Hagedorn, R. 2008. "Using the ECMWF Reforecast Dataset to Calibrate EPS Forecasts". *ECMWF Newsletter* 117 : 8-13. doi: 10.21957/TLK9S1TZZR.

Hagedorn, R., T.M. Hamill and J.S. Whitaker. 2008. "Probabilistic Forecast Calibration Using ECMWF and GFS Ensemble Reforecasts. Part I: Two-Meter Temperatures". *Monthly Weather Review* 136(7). American Meteorological Society: 2608–2619. doi: 10.1175/2007MWR2410.1.

Haiden, T., A. Kann, C. Wittmann, G. Pistotnik, B. Bica and C. Gruber. 2011. "The Integrated Nowcasting through Comprehensive Analysis (INCA) System and Its Validation over the Eastern Alpine Region". *Weather and Forecasting* 26(2). American Meteorological Society: 166–183. doi: 10.1175/2010WAF2222451.1.

Hamill, T.M. 2001. "Interpretation of Rank Histograms for Verifying Ensemble Forecasts". *Monthly Weather Review* 129(3). American Meteorological Society: 550–560. doi: 10.1175/1520-0493(2001)129<0550:IORHFV>2.0.CO;2.

Hamill, T.M. 2007. “Comments on ‘Calibrated Surface Temperature Forecasts from the Canadian Ensemble Prediction System Using Bayesian Model Averaging’”. *Monthly Weather Review* 135(12). American Meteorological Society: 4226–4230. doi: 10.1175/2007MWR1963.1.

Hamill, T.M. 2012. “Verification of TIGGE Multimodel and ECMWF Reforecast-Calibrated Probabilistic Precipitation Forecasts over the Contiguous United States”. *Monthly Weather Review* 140(7). American Meteorological Society: 2232–2252. doi: 10.1175/MWR-D-11-00220.1.

Hamill, T.M. 2018. “Practical Aspects of Statistical Postprocessing”. *Statistical Postprocessing of Ensemble Forecasts*. Elsevier, pp. 187–217. doi: 10.1016/B978-0-12-812372-0.00007-8.

Hamill, T.M. and M. Scheuerer. 2018. “Probabilistic Precipitation Forecast Postprocessing Using Quantile Mapping and Rank-Weighted Best-Member Dressing”. *Monthly Weather Review* 146(12). American Meteorological Society: 4079–4098. doi: 10.1175/MWR-D-18-0147.1. Also: Online appendix 1.

Hamill, T.M. and J.S. Whitaker. 2006. “Probabilistic Quantitative Precipitation Forecasts Based on Reforecast Analogs: Theory and Application”. *Monthly Weather Review* 134(11). American Meteorological Society: 3209–3229. doi: 10.1175/MWR3237.1.

Hamill, T.M., J.S. Whitaker and X. Wei. 2004. “Ensemble Reforecasting: Improving Medium-Range Forecast Skill Using Retrospective Forecasts”. *Monthly Weather Review* 132(6). American Meteorological Society: 1434–1447. doi: 10.1175/1520-0493(2004)132<1434:ERIMFS>2.0.CO;2.

Hamill, T.M., R. Hagedorn and J.S. Whitaker. 2008. “Probabilistic Forecast Calibration Using ECMWF and GFS Ensemble Reforecasts. Part II: Precipitation”. *Monthly Weather Review* 136(7). American Meteorological Society: 2620–2632. doi: 10.1175/2007MWR2411.1.

Hamill, T.M., G.T. Bates, J.S. Whitaker, D.R. Murray, M. Fiorino, T.J. Galarneau, Y. Zhu and W. Lapenta. 2013. “NOAA’s Second-Generation Global Medium-Range Ensemble Reforecast Dataset”. *Bulletin of the American Meteorological Society* 94(10). American Meteorological Society: 1553–1565. doi: 10.1175/BAMS-D-12-00014.1.

Hamill, T.M., E. Engle, D. Myrick, M. Peroutka, C. Finan and M. Scheuerer. 2017. “The U.S. National Blend of Models for Statistical Postprocessing of Probability of Precipitation and Deterministic Precipitation Amount”. *Monthly Weather Review* 145(9). American Meteorological Society: 3441–3463. doi: 10.1175/MWR-D-16-0331.1.

Hemri, S., D. Lisniak and B. Klein. 2015. “Multivariate Postprocessing Techniques for Probabilistic Hydrological Forecasting”. *Water Resources Research* 51(9): 7436–7451. doi: https://doi.org/10.1002/2014WR016473.

Hodyss, D., E. Satterfield, J. McLay, T.M. Hamill and M. Scheuerer. 2016. “Inaccuracies with Multimodel Postprocessing Methods Involving Weighted, Regression-Corrected Forecasts”. *Monthly Weather Review* 144(4). American Meteorological Society: 1649–1668. doi: 10.1175/MWR-D-15-0204.1.

Hoffman, R.N. and E. Kalnay. 1983. “Lagged Average Forecasting, an Alternative to Monte Carlo Forecasting”. *Tellus A* 35A(2): 100–118. doi: https://doi.org/10.3402/tellusa.v35i2.11425.

Hsieh, W. and B. Tang. 1998. “Applying Neural Network Models to Prediction and Data Analysis in Meteorology and Oceanography”. *Bulletin of The American Meteorological Society* 79: 1855–1870. doi: 10.1175/1520-0477(1998)079<1855:ANNMTP>2.0.CO;2.

Hu, Q., Z. Li, L. Wang, Y. Huang, Y. Wang and L. Li. 2019. “Rainfall Spatial Estimations: A Review from Spatial Interpolation to Multi-Source Data Merging”. *Water* 11(3). Multidisciplinary Digital Publishing Institute: 579. doi: 10.3390/w11030579.

Hwang, Y., A.J. Clark, V. Lakshmanan and S.E. Koch. 2015. “Improved Nowcasts by Blending Extrapolation and Model Forecasts”. *Weather and Forecasting* 30(5). American Meteorological Society: 1201–1217. doi: 10.1175/WAF-D-15-0057.1.

Jacks, E., J.B. Bower, V.J. Dagostaro, J.P. Dallavalle, M.C. Erickson and J.C. Su. 1990. “New NGM-Based MOS Guidance for Maximum/Minimum Temperature, Probability of Precipitation, Cloud Amount, and Surface Wind”. *Weather and Forecasting* 5(1). American Meteorological Society: 128–138. doi: 10.1175/1520-0434(1990)005<0128:NNBMGF>2.0.CO;2.

Jolliffe, I.T. and D.B. Stephenson, eds. 2012. *Forecast Verification: A Practitioner’s Guide in Atmospheric Science*. 2nd edition. Chichester. Wiley. https://onlinelibrary.wiley.com/doi/book/10.1002/9781119960003.

Jordan, A., F. Krueger and S. Lerch. 2020. “Scoring Rules for Parametric and Simulated Distribution Forecasts”. https://cran.r-project.org/package=scoringRules.

Kalman, R.E. 1960. “A New Approach to Linear Filtering and Prediction Problems”. *Journal of Basic Engineering* 82(1): 35–45. doi: 10.1115/1.3662552.

Keil, C. and G. Craig. 2011. “Regime-Dependent Forecast Uncertainty of Convective Precipitation”. *Meteorologische Zeitschrift* 20: 145–151. doi: 10.1127/0941-2948/2011/0219.

Kober, K., G.C. Craig, C. Keil and A. Dörnbrack. 2012. "Blending a Probabilistic Nowcasting Method with a High-Resolution Numerical Weather Prediction Ensemble for Convective Precipitation Forecasts". *Quarterly Journal of the Royal Meteorological Society* 138(664): 755–768. doi: https://doi.org/10.1002/qj.939.

Kober, K., G.C. Craig and C. Keil. 2014. "Aspects of Short-Term Probabilistic Blending in Different Weather Regimes". *Quarterly Journal of the Royal Meteorological Society* 140(681): 1179–1188. doi: https://doi.org/10.1002/qj.2220.

Lalaurette, F. 2003. "Early Detection of Abnormal Weather Conditions Using a Probabilistic Extreme Forecast Index". *Quarterly Journal of the Royal Meteorological Society* 129(594): 3037–3057. doi: https://doi.org/10.1256/qj.02.152.

Lin, H. 2012. "The Extreme Forecast Index (EFI) based on the Canadian Global Ensemble Prediction System (GEPS)". Environment and Climate Change Canada (formerly Environment Canada). https://collaboration.cmc.ec.gc.ca/cmc/cmoi/product_guide/docs/lib/op_systems/doc_opchanges/technote_geps-efi_20121003_e.pdf.

Lin, H., N. Gagnon, S. Beauregard, R. Muncaster, M. Markovic, B. Denis and M. Charron. 2016. "GEPS-Based Monthly Prediction at the Canadian Meteorological Centre". *Monthly Weather Review* 144(12). American Meteorological Society: 4867–4883. doi: 10.1175/MWR-D-16-0138.1.

Lorenz, E.N. 1969. "Atmospheric Predictability as Revealed by Naturally Occurring Analogues". *Journal of the Atmospheric Sciences* 26(4). American Meteorological Society: 636–646. doi: 10.1175/1520-0469(1969)26<636:APARBN>2.0.CO;2.

Mapes, B.E. 2000. "Convective Inhibition, Subgrid-Scale Triggering Energy, and Stratiform Instability in a Toy Tropical Wave Model". *Journal of the Atmospheric Sciences* 57(10). American Meteorological Society: 1515–1535. doi: 10.1175/1520-0469(2000)057<1515:CISSTE>2.0.CO;2.

Maraun, D. and M. Widmann. 2018. *Statistical Downscaling and Bias Correction for Climate Research*. Cambridge: Cambridge University Press. doi: 10.1017/9781107588783.

Markowski, P.M. and Y. Richardson. 2010. *Markowski, P: Mesoscale Meteorology in Midlatitudes*. Illustrated Edition. Chichester, UK. John Wiley & Sons, Inc.

Marty, R., V. Fortin, H. Kuswanto, A.-C. Favre and E. Parent. 2015. "Combining the Bayesian Processor of Output with Bayesian Model Averaging for Reliable Ensemble Forecasting". *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 64(1): 75–92. doi: https://doi.org/10.1111/rssc.12062.

Marzban, C. 2003. "Neural Networks for Postprocessing Model Output: ARPS". *Monthly Weather Review* 131(6). American Meteorological Society: 1103–1111. doi: 10.1175/1520-0493(2003)131<1103:NNFPMO>2.0.CO;2.

Masterton, J.M. and F.A. Richardson. 1979. *Humidex : A Method of Quantifying Human Discomfort Due to Excessive Heat and Humidity*. Environment Canada, Atmospheric Environment.

Messner, J. 2018. Ensemble Postprocessing With R, in *Statistical Postprocessing of Ensemble Forecasts*, eds. Vannitsem, S., D.S. Wilks and J. Messner, Elsevier, 291–329.

Metta, S., J. von Hardenberg, L. Ferraris, N. Rebora and A. Provenzale. 2009. "Precipitation Nowcasting by a Spectral-Based Nonlinear Stochastic Model". *Journal of Hydrometeorology* 10(5). American Meteorological Society: 1285–1297. doi: 10.1175/2009JHM1120.1.

Moisselin, J.-M., P. Cau, C. Jauffret, I. Bouissières and R. Tzanos. 2019. *Seamless approach for precipitations within the 0-3 hours forecast-interval*. 3rd European Nowcasting Conference. Agencia Estatal de Meteorología.

Nagarajan, B., L.D. Monache, J.P. Hacker, D.L. Rife, K. Searight, J.C. Knievel and T.N. Nipen. 2015. "An Evaluation of Analog-Based Postprocessing Methods across Several Variables and Forecast Models". *Weather and Forecasting* 30(6). American Meteorological Society: 1623–1643. doi: 10.1175/WAF-D-14-00081.1.

Nerini, D., L. Foresti, D. Leuenberger, S. Robert and U. Germann. 2019. "A Reduced-Space Ensemble Kalman Filter Approach for Flow-Dependent Integration of Radar Extrapolation Nowcasts and NWP Precipitation Ensembles". *Monthly Weather Review* 147(3). American Meteorological Society: 987–1006. doi: 10.1175/MWR-D-18-0258.1.

Nicolis, C. 1998. "Atmospheric Analogs and Recurrence Time Statistics: Toward a Dynamical Formulation". *Journal of the Atmospheric Sciences* 55(3). American Meteorological Society: 465–475. doi: 10.1175/1520-0469(1998)055<0465:AAARTS>2.0.CO;2.

Osczevski, R. and M. Bluestein. 2005. "The New Wind Chill Equivalent Temperature Chart". *Bulletin of the American Meteorological Society* 86(10). American Meteorological Society: 1453–1458. doi: 10.1175/BAMS-86-10-1453.

Perez-Zanon, N., L.-P. Caron, C. Alvarez-Castro, L. Batte, J. von Hardenberg, L. LLedo, N. Manubens, E. Sanchez-Garcia, B. van Schaeybroeck, V. Torralba and D. Verfaillie. 2021. CSTools: Assessing Skill of Climate Forecasts on Seasonal-to-Decadal Timescales. https://cran.r-project.org/package=CSTools.

Pierce, C.E., P.J. Hardaker, C.G. Collier and C.M. Haggett. 2000. "GANDOLF: A System for Generating Automated Nowcasts of Convective Precipitation". *Meteorological Applications* 7(4): 341–360. doi: https://doi.org/10.1017/S135048270000164X.

Pillosu, F. and T. Hewson. 2017. "New point-rainfall forecasts for flash flood prediction". *ECMWF Newsletter* 153. https://www.ecmwf.int/en/newsletter/153/news/new-point-rainfall-forecasts-flash-flood-prediction.

Poletti, M.L., F. Silvestro, S. Davolio, F. Pignone and N. Rebora. 2019. "Using Nowcasting Technique and Data Assimilation in a Meteorological Model to Improve Very Short Range Hydrological Forecasts". *Hydrology and Earth System Sciences* 23: 3823–3841. doi: 10.5194/hess-23-3823-2019.

Radhakrishna, B., I. Zawadzki and F. Fabry. 2012. "Predictability of Precipitation from Continental Radar Images. Part V: Growth and Decay". *Journal of the Atmospheric Sciences* 69(11). American Meteorological Society: 3336–3349. doi: 10.1175/JAS-D-12-029.1.

Raftery, A.E., T. Gneiting, F. Balabdaoui and M. Polakowski. 2005. "Using Bayesian Model Averaging to Calibrate Forecast Ensembles". *Monthly Weather Review* 133(5). American Meteorological Society: 1155–1174. doi: 10.1175/MWR2906.1.

Rasp, S. and S. Lerch. 2018. "Neural Networks for Postprocessing Ensemble Weather Forecasts". *Monthly Weather Review* 146(11). American Meteorological Society: 3885–3900. doi: 10.1175/MWR-D-18-0187.1.

Raynaud, L., O. Pannekoucke, P. Arbogast and F. Bouttier. 2015. "Application of a Bayesian Weighting for Short-Range Lagged Ensemble Forecasting at the Convective Scale". *Quarterly Journal of the Royal Meteorological Society* 141(687): 459–468. doi: https://doi.org/10.1002/qj.2366.

Roulston, M.S. and L.A. Smith. 2003. "Combining Dynamical and Statistical Ensembles". *Tellus A: Dynamic Meteorology and Oceanography* 55(1). Taylor & Francis: 16–30. doi: 10.3402/tellusa.v55i1.12082.

Schefzik R. and A. Möller. 2018. Ensemble methods incorporating dependence structures, in *Statistical Postprocessing of Ensemble Forecasts*, eds. Vannitsem, S., D.S. Wilks and J. Messner, Elsevier, 91–125.

Schefzik, R. 2017. "Ensemble Calibration with Preserved Correlations: Unifying and Comparing Ensemble Copula Coupling and Member-by-Member Postprocessing". *Quarterly Journal of the Royal Meteorological Society* 143(703): 999–1008. doi: https://doi.org/10.1002/qj.2984.

Scheuerer, M. and T. M. Hamill. 2015. "Statistical Postprocessing of Ensemble Precipitation Forecasts by Fitting Censored, Shifted Gamma Distributions". *Monthly Weather Review* 143(11). American Meteorological Society: 4578–4596. doi: https://doi.org/10.1175/MWR-D-15-0061.1.

Scheuerer, M. and T.M. Hamill. 2019. "Probabilistic Forecasting of Snowfall Amounts Using a Hybrid between a Parametric and an Analog Approach". *Monthly Weather Review* 147(3). American Meteorological Society: 1047–1064. doi: 10.1175/MWR-D-18-0273.1.

Scheuerer, M., S. Gregory, T.M. Hamill and P.E. Shafer. 2017. "Probabilistic Precipitation-Type Forecasting Based on GEFS Ensemble Forecasts of Vertical Temperature Profiles". *Monthly Weather Review* 145(4). American Meteorological Society: 1401–1412. doi: 10.1175/MWR-D-16-0321.1.

Scheufele, K., K. Kober, G.C. Craig and C. Keil. 2014. "Combining Probabilistic Precipitation Forecasts from a Nowcasting Technique with a Time-Lagged Ensemble". *Meteorological Applications* 21(2): 230–240. doi: https://doi.org/10.1002/met.1381.

Schmid, W., S. Mecklenburg and J. Joss. 2000. "Short-Term Risk Forecasts of Severe Weather". *Physics and Chemistry of the Earth, Part B: Hydrology, Oceans and Atmosphere* 25(10). First European Conference on Radar Meteorology: 1335–1338. doi: 10.1016/S1464-1909(00)00204-5.

Schuhen, N., T.L. Thorarinsdottir and T. Gneiting. 2012. "Ensemble Model Output Statistics for Wind Vectors". *Monthly Weather Review* 140(10). American Meteorological Society: 3204–3219. doi: 10.1175/MWR-D-12-00028.1.

Seed, A.W. 2003. "A Dynamic and Spatial Scaling Approach to Advection Forecasting". *Journal of Applied Meteorology and Climatology* 42(3). American Meteorological Society: 381–388. doi: 10.1175/1520-0450(2003)042<0381:ADASSA>2.0.CO;2.

Showalter, A.K. 1953. "A Stability Index for Thunderstorm Forecasting". *Bulletin of the American Meteorological Society* 34(6). American Meteorological Society: 250–252. doi: 10.1175/1520-0477-34.6.250.

Sideris, I.V., L. Foresti, D. Nerini and U. Germann. 2020. "NowPrecip: Localized Precipitation Nowcasting in the Complex Terrain of Switzerland". *Quarterly Journal of the Royal Meteorological Society* 146(729): 1768–1800. doi: https://doi.org/10.1002/qj.3766.

Siegert, S., J. Bhend, I. Kroener and M. De Felice. 2020. SpecsVerification: Forecast Verification Routines for Ensemble Forecasts of Weather and Climate. https://cran.r-project.org/web/packages/SpecsVerification/index.html.

Simonin, D., C. Pierce, N. Roberts, S.P. Ballard and Z. Li. 2017. "Performance of Met Office Hourly Cycling NWP-Based Nowcasting for Precipitation Forecasts". *Quarterly Journal of the Royal Meteorological Society* 143(708): 2862–2873. doi: https://doi.org/10.1002/qj.3136.

Sloughter, J.M., T. Gneiting and A.E. Raftery. 2010. "Probabilistic Wind Speed Forecasting Using Ensembles and Bayesian Model Averaging". *Journal of the American Statistical Association* 105(489). Taylor & Francis: 25–35. doi: 10.1198/jasa.2009.ap08615.

Stensrud, D.J. and M.S. Wandishin. 2000. "The Correspondence Ratio in Forecast Evaluation". *Weather and Forecasting* 15(5). American Meteorological Society: 593–602. doi: 10.1175/1520-0434(2000)015<0593:TCRIFE>2.0.CO;2.

Taillardat, M., O. Mestre, M. Zamo and P. Naveau. 2016. "Calibrated Ensemble Forecasts Using Quantile Regression Forests and Ensemble Model Output Statistics". *Monthly Weather Review* 144(6). American Meteorological Society: 2375–2393. doi: 10.1175/MWR-D-15-0260.1.

Thompson, R.L., C.M. Mead and R. Edwards. 2007. "Effective Storm-Relative Helicity and Bulk Shear in Supercell Thunderstorm Environments". *Weather and Forecasting* 22(1). American Meteorological Society: 102–115. doi: 10.1175/WAF969.1.

Toth, Z., O. Talagrand and Y. Zhu. 2006. The Attributes of Forecast Systems: A General Framework for the Evaluation and Calibration of Weather Forecasts, in *Predictability of Weather and Climate*, eds. Hagedorn, R. and T. Palmer, Cambridge University Press, pp. 584–595. doi: 10.1017/CBO9780511617652.023.

Tsonis, A.A. and G.L. Austin. 1981. "An Evaluation of Extrapolation Techniques for the Short-term Prediction of Rain Amounts". *Atmosphere-Ocean* 19: 54–65. doi: 10.1080/07055900.1981.9649100.

Van Schaeybroeck, B. and S. Vannitsem. 2011. "Post-Processing through Linear Regression". *Nonlinear Processes in Geophysics* 18(2). Copernicus GmbH: 147–160. doi: 10.5194/npg-18-147-2011.

Van Schaeybroeck, B. and S. Vannitsem. 2015. "Ensemble Post-Processing Using Member-by-Member Approaches: Theoretical Aspects". *Quarterly Journal of the Royal Meteorological Society* 141(688): 807–818. doi: https://doi.org/10.1002/qj.2397.

Vannitsem, S. 2009. "A Unified Linear Model Output Statistics Scheme for Both Deterministic and Ensemble Forecasts". *Quarterly Journal of the Royal Meteorological Society* 135(644): 1801–1815. doi: https://doi.org/10.1002/qj.491.

Vannitsem, S. and R. Hagedorn. 2011. "Ensemble Forecast Post-Processing over Belgium: Comparison of Deterministic-like and Ensemble Regression Methods". *Meteorological Applications* 18(1): 94–104. doi: https://doi.org/10.1002/met.217.

Vannitsem, S., D.S. Wilks and J. Messner, eds. 2018a. *Statistical Postprocessing of Ensemble Forecasts*. 1st edition. Cambridge, MA. Elsevier.

Vislocky, R.L. and J.M. Fritsch. 1995. "Improved Model Output Statistics Forecasts through Model Consensus". *Bulletin of the American Meteorological Society* 76: 1157–1164. doi: 10.1175/1520-0477(1995)076<1157:IMOSFT>2.0.CO;2.

Vitart, F., G. Balsamo, J.-R. Bidlot, S. Lang, I. Tsonevsky, D. Richardson and M. Alonso-Balmaseda. 2019. "Use of ERA5 to Initialize Ensemble Re-Forecasts". ECMWF. doi: 10.21957/w8i57wuz6.

Whitaker, J.S., X. Wei and F. Vitart. 2006. "Improving Week-2 Forecasts with Multimodel Reforecast Ensembles". *Monthly Weather Review* 134(8). American Meteorological Society: 2279–2284. doi: 10.1175/MWR3175.1.

Wilks, D.S. 2018. Univariate ensemble, in *Statistical Postprocessing of Ensemble Forecasts*, eds. Vannitsem, S., D.S. Wilks and J. Messner, Elsevier, 49-89.

Wilks, D.S. 2019. *Statistical Methods in the Atmospheric Sciences*. Fourth edition. Amsterdam, Elsevier. ISBN 9780128158234, eBook ISBN 9780128165270.

Wilson, L.J., S. Beauregard, A.E. Raftery and R. Verret. 2007. "Calibrated Surface Temperature Forecasts from the Canadian Ensemble Prediction System Using Bayesian Model Averaging". *Monthly Weather Review* 135(4). American Meteorological Society: 1364–1385. doi: 10.1175/MWR3347.1.

Yuen, R.A., S. Baran, C. Fraley, T. Gneiting, S. Lerch, M. Scheuerer and T. Thorarinsdottir. 2018. ensembleMOS: Ensemble Model Output Statistics. https://cran.r-project.org/web/packages/ensembleMOS/index.html.

Zhu, Y. and Y. Luo. 2015. "Precipitation Calibration Based on the Frequency-Matching Method". *Weather and Forecasting* 30(5). American Meteorological Society: 1109–1124. doi: 10.1175/WAF-D-13-00049.1.

Zimmer, M., G.C. Craig, C. Keil and H. Wernli. 2011. "Classification of Precipitation Events with a Convective Response Timescale and Their Forecasting Characteristics". *Geophysical Research Letters* 38. doi: 10.1029/2010GL046199.

Zsoter, E. 2006. "Recent developments in extreme weather forecasting". *ECMWF Newsletter* 107: 8-17. doi: 10.21957/kl9821hnc7.

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 87 40/83 14 – Fax: +41 (0) 22 730 80 27

Email: communications@wmo.int

public.wmo.int

JN 20579