BAMS
Essay

Statistical Postprocessing for Weather Forecasts

Review, Challenges, and Avenues in a Big Data World

Stéphane Vannitsem, John Bjørnar Bremnes, Jonathan Demaeyer, Gavin R. Evans, Jonathan Flowerdew, Stephan Hemri, Sebastian Lerch, Nigel Roberts, Susanne Theis, Aitor Atencia, Zied Ben Bouallègue, Jonas Bhend, Markus Dabernig, Lesley De Cruz, Leila Hieta, Olivier Mestre, Lionel Moret, Iris Odak Plenković, Maurice Schmeits, Maxime Taillardat, Joris Van den Bergh, Bert Van Schaeybroeck, Kirien Whan, and Jussi Ylhaisi

ABSTRACT: Statistical postprocessing techniques are nowadays key components of the forecasting suites in many national meteorological services (NMS), with, for most of them, the objective of correcting the impact of different types of errors on the forecasts. The final aim is to provide optimal, automated, seamless forecasts for end users. Many techniques are now flourishing in the statistical, meteorological, climatological, hydrological, and engineering communities. The methods range in complexity from simple bias corrections to very sophisticated distribution-adjusting techniques that incorporate correlations among the prognostic variables. The paper is an attempt to summarize the main activities going on in this area from theoretical developments to operational applications, with a focus on the current challenges and potential avenues in the field. Among these challenges is the shift in NMS toward running ensemble numerical weather prediction (NWP) systems at the kilometer scale that produce very large datasets and require high-density high-quality observations, the necessity to preserve space–time correlation of high-dimensional corrected fields, the need to reduce the impact of model changes affecting the parameters of the corrections, the necessity for techniques to merge different types of forecasts and ensembles with different behaviors, and finally the ability to transfer research on statistical postprocessing to operations. Potential new avenues are also discussed.

KEYWORDS: Bias; Operational forecasting; Probability forecasts/models/distribution; Model output statistics; Data science; Regression

https://doi.org/10.1175/BAMS-D-19-0308.1
Corresponding author: Stéphane Vannitsem, stephane.vannitsem@meteo.be
In final form 28 August 2020

AFFILIATIONS: Vannitsem and Demaeyer—Royal Meteorological Institute of Belgium, and European Meteorological Network (EUMETNET), Brussels, Belgium; **Bremnes**—Norwegian Meteorological Institute, Oslo, Norway; **Evans and Flowerdew**—Met Office, Exeter, United Kingdom; **Hemri**—Federal Office of Meteorology and Climatology, MeteoSwiss, Zurich, Switzerland; **Lerch**—Institute for Stochastics, Karlsruhe Institute of Technology, Karlsruhe, Germany; **Roberts**—MetOffice@Reading, Met Office, United Kingdom; **Theis**—Deutscher Wetterdienst, Offenbach, Germany; **Atencia and Dabernig**—Zentralanstalt für Meteorologie und Geodynamik, Vienna, Austria; **Ben Bouallègue**—European Centre for Medium-Range Weather Forecasts, Reading, United Kingdom; **Bhend and Moret**—Federal Office of Meteorology and Climatology, MeteoSwiss, Zurich, Switzerland; **De Cruz, Van den Bergh, and Van Schaeybroeck**—Royal Meteorological Institute of Belgium, Brussels, Belgium; **Hieta and Ylhaisi**—Finnish Meteorological Institute, Helsinki, Finland; **Mestre and Taillardat**—Météo-France, CNRM-UMR 3589, Toulouse, France; **Odak Plenković**—Croatian Meteorological and Hydrological Service, Zagreb, Croatia; **Schmeits and Whan**—Royal Netherlands Meteorological Institute (KNMI), De Bilt, The Netherlands

Errors from multiple sources have a detrimental effect on the skill of weather forecasts. One of the primary sources of errors is associated with the construction of an initial condition for numerical weather forecasting systems. These errors grow rapidly during the course of the forecasts until they reach a level beyond which the forecasts do not display any useful skill. This property is known as the sensitivity to initial conditions; see, e.g., the review of Vannitsem (2017). Two other important categories of errors that reduce forecast skill are the boundary-condition errors (e.g., Collins and Allen 2002; Nicolis 2007) and the model structural errors. Model structural errors include a missing or poor representation of subgrid dynamical and physical processes and inaccuracies associated with the numerical scheme (Lorenz 1982; Nicolis et al. 2009). All these numerical weather prediction (NWP) model deficiencies induce errors that are rapidly amplified in time due to the chaotic nature of the model dynamics, and in turn affect the course of the forecasts by inducing what are often called systematic and random errors.

Since the early 1950s, numerical weather systems have been developed with an ever-increasing complexity (Lynch 2008). In conjunction, the quality of the forecasts has been constantly improving with current forecasts deemed useful even beyond 15 days (Buizza and Leutbecher 2015; Bauer et al. 2015; Van Straaten et al. 2020). This success can be attributed to both improvements in the quality of the initial conditions of the numerical prediction models coming from improved data assimilation, and improvements in the model representation of physical and dynamical processes; or in other words a reduction of both initial condition and model errors. Nowadays, many operational forecasting centers are running both single high-resolution deterministic forecasts (at kilometer-scale grid spacing) and ensemble forecasts (often at kilometer scale too), the latter providing information about the probability of occurrence of specific atmospheric states (Toth and Kalnay 1993; Molteni et al. 1996; Yoden 2007; Leutbecher and Palmer 2008; Buizza 2019; Frogner et al. 2019; Schwartz et al. 2019).

Despite these major developments, both deterministic and ensemble forecasts continue to display significant deficiencies associated with the presence of model errors and an inappropriate distribution of initial conditions. This results in systematic biases and inappropriate dispersion of ensemble forecasts requiring some sort of postprocessing in order to improve the forecast quality. Statistical postprocessing methods used for this purpose involve a wide range of correction techniques that can be appropriately developed for either deterministic or ensemble forecasts (Wilks 2011; Vannitsem et al. 2018).

The first applications and operational implementations of statistical corrections were based on simple linear regression techniques using linear statistical relations deduced

from observational data only, known as perfect prog, or built between the observations and predictors generated by the weather forecasting models, known as model output statistics (Klein et al. 1959; Glahn and Lowry 1972). These approaches were successful, attracting much interest from the meteorological community with many applications to a wide range of model variables and extensions to other types of regression functions like logistic regression or neural networks (e.g., Lemcke and Kruizinga 1988; Marzban 2003). Nowadays there is a bloom of techniques, in particular, for ensemble forecasts with the purpose of producing probabilistic information with a more accurate representation of forecast uncertainty (Gneiting et al. 2007).

Since postprocessing aims to improve the quality and usefulness of the forecasts, an important aspect is the choice of measures used to assess that quality. Three key attributes of ensemble forecasts are usually sought: first, the forecasts should be as close as possible to the truth or the observations (the proxy for the truth), given the constraint of the underlying uncertainty; second, the forecast should respect the climatology or the frequency with which different thresholds are exceeded, a climatological reliability; and third, the observation should be statistically indistinguishable from the forecasts produced by the model, assuming the representativeness of the observation point is also included (Gneiting and Raftery 2007; Wilks 2011). The first property is related to the resolution/sharpness of the forecasts and the two latter to the reliability. Statistical calibration provides a natural way to improve reliability, essential for rational decision making, sometimes at the expense of resolution/sharpness. Verification can help to identify the key systematic errors which the postprocessing should be designed to address, check its success in correcting them, and the impact on wider performance measures. A variety of scoring rules and diagnostic tools exists for this purpose (e.g., Richardson 2000; Wilks 2011; Thorarinsdottir and Schuhen 2018). For ensemble forecasts, Brier skill scores, rank probability skill scores, rank histograms, or spread–skill relationships are popular measures.

Statistical correction techniques for both deterministic and ensemble forecasts should nowadays be an integral part of any operational forecasting system. As illustrated in Hemri et al. (2014), whatever the degree of sophistication of the model under consideration, the statistical postprocessing approach still provides additional corrections that will benefit the end user. This last consideration should make research into this area and operational implementation key priorities of national meteorological services (NMS).

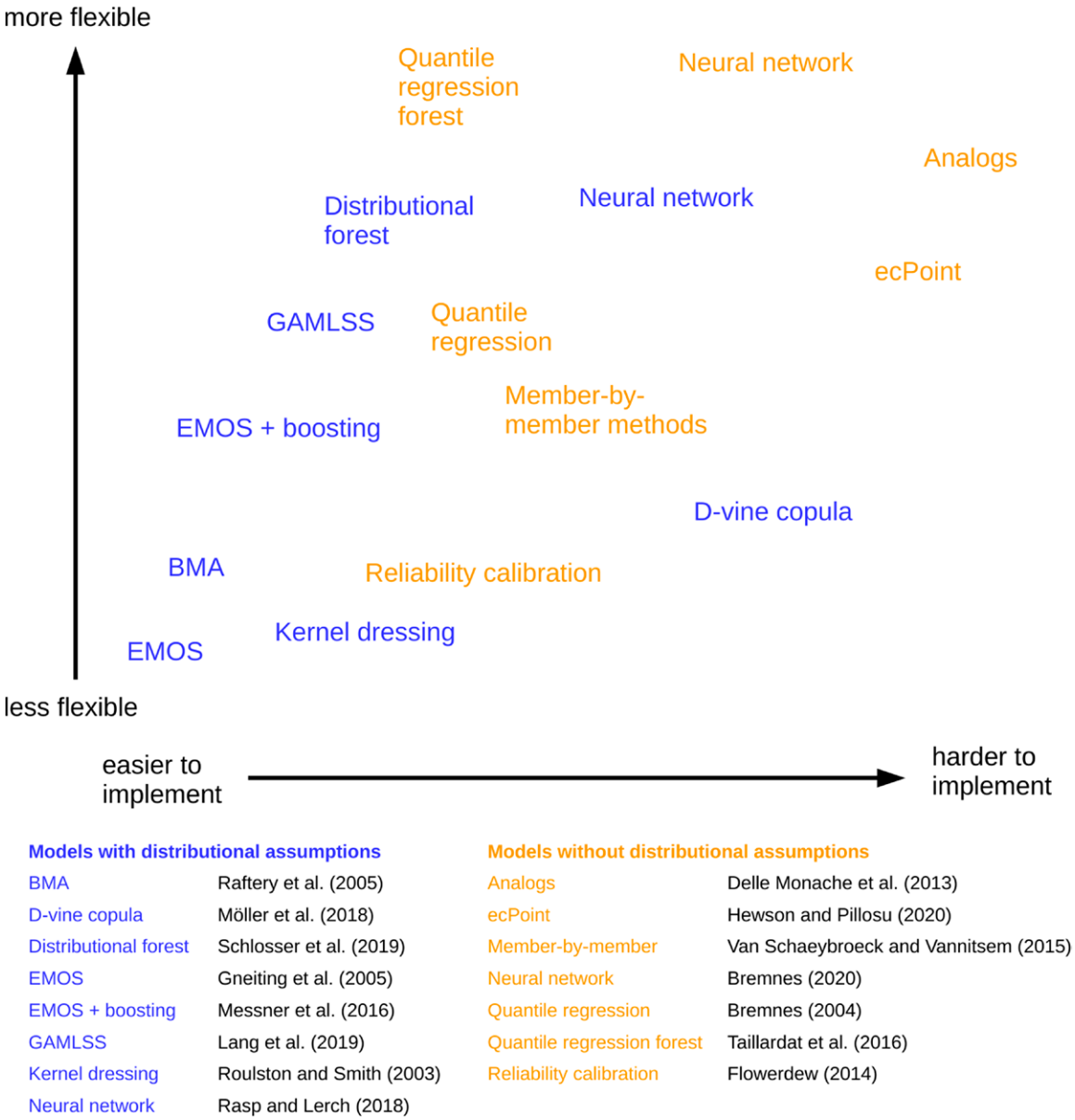
The goal of this paper is to review the current research developments and operational implementations taking place worldwide and particularly in Europe, together with the future prospects and challenges in the area of statistical postprocessing. A key challenge is the exponential growth of data that are available from both the model forecasts and the observations, accompanied by an ever increasing need for very localized, yet seamless, forecast information.

The second section discusses state-of-the-art approaches for statistical postprocessing of ensemble forecasts. The impact of using statistical postprocessing on the physical coherence of the dynamics is addressed in the third section. The fourth section covers the problem of the frequent model changes that could affect the quality of the statistical correction techniques. In the fifth section, the use of blending techniques for correcting the forecasts and providing seamless probabilistic information is reviewed. The sixth section evaluates the potential implementation difficulties of the statistical correction techniques. Finally, future prospects and challenges are discussed in the seventh section.

State-of-the-art statistical postprocessing methods

From a statistical perspective, most postprocessing methods can be categorized into two groups—those that assume the predictive distribution belongs to a class of known probability distributions and those that do not. Historically, these groups are often respectively referred

to as "parametric" and "nonparametric" approaches. In the present work, we will refer to these two classes of methodologies as "distribution-based assumptions" and "distribution-free assumptions" approaches depending on whether a distribution is considered a priori or not. Recent developments for these two categories will be addressed separately, and then some available tools for both are listed. Some key methodological challenges are briefly discussed at the end of the section. Figure 1 illustrates a set of these techniques in a two dimensional representation of the flexibility with respect to their applicability versus the requirements for



Models with distributional assumptions		Models without distributional assumptions	
BMA	Raftery et al. (2005)	Analogs	Delle Monache et al. (2013)
D-vine copula	Möller et al. (2018)	ecPoint	Hewson and Pillosu (2020)
Distributional forest	Schlosser et al. (2019)	Member-by-member	Van Schaeybroeck and Vannitsem (2015)
EMOS	Gneiting et al. (2005)	Neural network	Bremnes (2020)
EMOS + boosting	Messner et al. (2016)	Quantile regression	Bremnes (2004)
GAMLSS	Lang et al. (2019)	Quantile regression forest	Taillardat et al. (2016)
Kernel dressing	Roulston and Smith (2003)	Reliability calibration	Flowerdew (2014)
Neural network	Rasp and Lerch (2018)		

Fig. 1. Overview of probabilistic postprocessing methods, with approaches with distributional assumptions shown in blue and approaches without distributional assumptions shown in orange. The horizontal axis refers to requirements in terms of training data, tuning parameter choices, available software implementations (also related to the novelty of the approaches), and model size. The vertical axis characterizes model flexibility and adaptability in terms of the applicability to different output variables, and the complexity of representable relationships between inputs (including multiple predictors) and output. As a matter of caution, this graph based on our current evaluation of the methods available should be viewed dynamically, and some techniques can rapidly move toward another corner of the graph. One example is the ecPoint method, which is a very new and promising concept, and if there is increased community experience in using such an approach, this method could move toward the left, and possibly upward, on the graph.

implementation. This is a subjective view based on the expertise of the authors only, which provides some guidelines for those starting in the field.

Development of approaches with distribution-based assumptions. Distribution-based postprocessing approaches specify a parametric model for the forecast distribution by selecting a suitable family of probability distributions depending on the variable of interest (Gneiting et al. 2005; Raftery et al. 2005). The parameters of the forecast distributions are then linked to the predictors from the NWP system via regression equations to correct systematic errors. The regression coefficients are estimated by optimizing suitable loss functions for distribution forecasts such as the continuous ranked probability score (CRPS; Gneiting and Raftery 2007). Such constructions lead to models that are straightforward to fit and are widely used at NMS. Summary statistics of the ensemble predictions of the variable of interest are often used as sole covariates. However, many more potential predictors (including predictions of other variables as well as geographical or temporal information) are usually available and might provide great benefit, but specifying their functional relations to the distribution parameters is challenging. One promising approach is gradient boosting for distributional regression (Messner et al. 2017) which selects the most important predictor variables during parameter estimation by iteratively updating the regression coefficient of the predictor that improves the current model fit the most.

Several variants where the forecast distribution remains parametric, but fewer assumptions concerning the functional relation between predictors and distribution parameters are required, have been proposed recently. Lang et al. (2019) and Simon et al. (2019) use generalized additive models where the parameters of a forecast distribution from a wide range of distributional families with parameterizations corresponding to location, scale, and shape are flexibly modeled as additive functions of the predictors (Rigby and Stasinopoulos 2005). The functional dependencies between distribution parameters and predictors are usually prescribed separately for each parameter. Schlosser et al. (2019) extend this framework by using regression trees and random forests to recursively partition the predictor space based on maximum likelihood estimation, resulting in separate distribution models for each partition. Rasp and Lerch (2018) propose the use of neural networks to link distribution parameters to predictors. Since neural networks are very flexible functions composed of a sequence of nonlinear transformations, they facilitate modeling exceedingly flexible relations jointly for all distribution parameters. By estimating the neural network weights/coefficients using optimization based on stochastic gradient descent algorithms, highly complex models can be fitted.

Despite these powerful variants the need to select a suitable parametric family to describe the distribution of the target variable remains a limitation for parametric postprocessing methods and often necessitates elaborate fine-tuning (Gebetsberger et al. 2017) or complex mixture models (Baran and Lerch 2016) to achieve well-calibrated forecast distributions. A flexible alternative to the full-distribution adjusted parametric methods just discussed above is the kernel dressing method, consisting of replacing individual ensemble members by kernel functions (Roulston and Smith 2003; Bröcker and Smith 2008). This approach can accommodate any type of ensemble continuous distribution, but can be considered also as a distribution-based approach since the parameters of the kernel need to be fitted.

Development of approaches with distribution-free assumptions. An alternative is to use postprocessing methods that avoid distributional assumptions by constructing approximations of the forecast distribution. Bremnes (2004) proposes the use of quantile regression methods in which only a selected set of quantiles of the predictive distribution is considered. The quantiles

are traditionally estimated separately which can lead to invalid crossing quantiles. However, by adding constraints to the estimation or by simply reordering the quantiles at the end, this problem can be circumvented. Quantile regression methods have recently been extended in several directions. Wahl (2015) proposes a penalization method in a Bayesian context where regularization and predictor selection are performed simultaneously. Ben Bouallègue (2017) suggests a similar penalized estimation approach that has proved useful in situations with many predictors. In Bremnes (2019) constrained splines are applied to make use of information from all ensemble members. Standard quantile regression methods are not suitable for extreme quantile levels and were the motivation of Velthoen et al. (2019) for developing an estimator for extreme levels based on local quantile regressions. Möller et al. (2018) apply D-vine copula regression methods to model the joint distribution of the weather variable of interest and the predictors, and derive the forecast distribution from this.

Cannon (2018) and Bremnes (2020) shift the focus from a finite set of quantiles to complete quantile functions, thus providing models of the entire forecast distribution. The former adds the quantile level as a monotone predictor in a neural network approach, while the latter assumes the quantile function to be a Bernstein polynomial and models the relation between its coefficients and the predictors by means of a neural network like Rasp and Lerch (2018). Similarly targeting the full predictive distribution, Veldkamp et al. (2021) apply a discretization to the target variable and model the predictive density by a histogram, with probabilities obtained as output of a convolutional neural network. Isotonic distributional regression, a nonparametric technique that takes advantage of partial order relations among the predictors and requires minimal implementation decisions only was recently proposed by Henzi et al. (2019).

Taillardat et al. (2016) propose a postprocessing model using quantile regression forests, a quantile regression method where predictive quantiles are computed based on random forests (Breiman 2001; Meinshausen 2006). Random forest methods have been used in a variety of postprocessing applications (Gagne et al. 2009, 2017; McGovern et al. 2017) and quantile regression forests have been applied to a wide range of weather variables (Taillardat et al. 2016; Zamo 2016; van Straaten et al. 2018; Whan and Schmeits 2018). Recent extensions include combinations with parametric distribution models fitted to forest-based outputs to circumvent the restriction of predictive quantiles by the range of training observations and to provide better forecasts for extreme events (Taillardat et al. 2019), as well as quantile regression forests calibration based on forecast anomalies to improve predictions of cold and heat waves (Taillardat and Mestre 2020). The now operational ecPoint methodology (Pillosu and Hewson 2017; Hewson and Pillosu 2020) utilizes a decision-tree method based on expert elicitation to derive, for the whole world, pointwise precipitation forecasts from gridbox forecasts by accounting for both the expected gridscale NWP bias and the expected subgrid variability, according to the diagnosed gridbox weather type.

Further, postprocessing methods based on historical analogs do not use any distributional assumptions and, therefore, can be viewed in a similar framework (Taillardat et al. 2016). Instead of obtaining sets of analogous historical cases by recursively partitioning the predictor space via random forests, analog-based methods (Hamill and Whitaker 2006; Delle Monache et al. 2011, 2013) sequentially search for the most similar past cases for every new input vector of predictor values using a specifically tailored similarity measure, usually based on weighted combinations of several predictors. Forecast distributions are then constructed from the corresponding set of past observations. Analog-based methods have been applied for a variety of variables (Alessandrini et al. 2015, 2018; Nagarajan et al. 2015; Odak Plenković et al. 2018, 2020). As for all postprocessing methods, there is a clear benefit of increasingly large training datasets: in this case, the benefit is that closer analogs can be found. However, the computational cost of determining suitable analogs may become prohibitive for large datasets, particularly in the presence of many possible covariates. Several

studies have proposed the use of analogs or related concepts to determine training datasets consisting of similar past forecast cases for methods with distributional assumptions (Hamill et al. 2008; Junk et al. 2015; Lerch and Baran 2017; Scheuerer and Hamill 2019; Schlosser et al. 2019) and or without (Bremnes 2004; Hamill et al. 2015). These approaches are motivated by the premise that customized training datasets consisting of similar past forecast cases may allow the use of simpler models that are better adapted to the current conditions, and are closely related to local modeling methods in statistics (e.g., Bottou and Vapnik 1992; Loader 1999).

Other distribution-free postprocessing methods use various direct transformations of ensembles, emerging from simple bias correction or variance inflation procedures (e.g., Johnson and Swinbank 2009). Flowerdew (2014) directly maps forecast probability to observed event frequency for a series of thresholds, to make the forecasts reliable (and thus unbiased and correctly spread) while preserving the resolution. Each threshold is calibrated using the most local region consistent with achieving a specified sample size, and the algorithm only needs to see each piece of training data once, reducing storage requirements for large gridded datasets. Approaches acting simultaneously on each member in an ensemble have been suggested by Van Schaeybroeck and Vannitsem (2015) and others. In these, all parameters defining the transformation of the ensemble are estimated jointly by either optimizing the empirical CRPS or following maximum likelihood principles under further distributional assumptions. Some available tools are listed in the appendix.

New methodological challenges. Most new methodologies are based around machine learning (ML) techniques. Although first attempts at using modern ML approaches for postprocessing have shown promising improvements over traditional approaches, a number of challenges remain. It is important that inherent structures in NWP forecasts (and in NWP errors) should be more fully exploited. For example, current approaches often rely on a limited set of ensemble statistics and covariates only, but modern ML approaches, in particular approaches involving neural networks, could be designed to target relationships between a large number of covariates more directly. These approaches can also be efficiently implemented on massively parallel supercomputers (e.g., Cervone et al. 2017). Furthermore, recent developments in deep learning, notably convolutional neural networks, have made it possible to use large gridded input datasets, allowing one to use more spatial information in statistical postprocessing; see Scheuerer et al. (2020) and Veldkamp et al. (2021) for first applications.

A second set of challenges relates to the interpretability of ML approaches. While many methods are regarded as "black boxes" various techniques can provide an understanding of what ML models have learned [see McGovern et al. (2019) for a recent review], for example, which predictors are most important in the model and for a particular forecast. Many approaches report global variable importance (Taillardat et al. 2016; Rasp and Lerch 2018; van Straaten et al. 2018; Whan and Schmeits 2018) that can be used as a sanity check for the model and can give insights about relationships between the response and the set of predictors. An explanation of individual predictions is of interest to many users and can be achieved with methods such as Shapley values (Molnar 2019). Due to their complexity, modern ML approaches, and deep neural networks in particular, can produce unexpected results at times due to overfitting or unstable output when slight modifications of the input are introduced (Antun et al. 2020).

Preserving space and time correlation

Weather forecasts typically include spatial, temporal, and multivariable information. Forecasters, and even the public, are nowadays accustomed to animated maps of multiple meteorological variables. Such visualizations reveal common patterns in space and time and between variables. These range from trivial relations (e.g., clustering of rainfall cells) to more complex relations (e.g.,

between radiation, temperature, and humidity), and together are called the dependence structure. Given the assumption that a modern NWP model can reproduce the correct dependence structure, the question is whether postprocessing methods are able to do so as well and hence adjust variables in a consistent way. This is also important in the context of downstream applications such as hydrological ensemble forecasting systems and renewable energy applications (Cloke and Pappenberger 2009; Pinson and Messner 2018), for which space–time coherence is crucial for realistic scenarios. For weather variables with continuous probability distributions such as temperature or wind, the dependence structure is automatically preserved by the so-called member-by-member postprocessing (MBMP) approaches (Doblas-Reyes et al. 2005; Johnson and Bowler 2009; Flowerdew 2014; Van Schaeybroeck and Vannitsem 2011, 2015). In contrast, techniques that use predictive distributions require additional approaches to reestablish the dependence structure (Schefzik and Möller 2018). One approach is based on parametric methods and makes use of specific multivariate predictive distributions that are suitable for low-dimensional settings or settings with specific intervariable relations (Pinson and Girard 2012). Higher-dimensional situations can be adequately handled by nonparametric methods such as ensemble copula coupling (ECC; Schefzik et al. 2013) and the Schaake shuffle (SSH; Clark et al. 2004; Sperati et al. 2017). ECC is based on empirical copulas aimed at restoring the dependence structure of the forecast and is derived from the rank order of the members in the raw ensemble forecast, under a perfect model assumption, with exchangeable ensemble members. For SSH, on the other hand, the dependence structure is derived from historical observations instead. Finally, dependencies can also be taken into account through univariate postprocessing methods with location-specific model parameters. These methods include geostatistical model averaging (Kleiber et al. 2011) and spatially adaptive models that make use of anomalies (Scheuerer and König 2014) or standardized anomalies as in standardized anomaly model output statistics (Dabernig et al. 2017, 2020). Figure 2 illustrates two specific routes of postprocessing that allow the preservation of intervariable dependencies for high-dimensional problems.

Despite their simplicity and efficiency, multivariate approaches like ECC or SSH are prone to introduce physically unrealistic artifacts into the forecast scenarios. SSH is not flow dependent and ECC is impaired by small spread and errors in the dependence structure of the raw ensemble (Schefzik and Möller 2018). Flow-dependent variants of the SSH that select historical observations for the dependence template based on some similarity measures have been developed by Schefzik (2016), Scheuerer et al. (2017), and Scheuerer and Hamill (2018). Ben Bouallègue et al. (2016) propose dual ECC as an ECC variant which corrects for errors in the dependence template by also considering the autocorrelation of past forecast errors. Furthermore, if the dependence template, i.e., the raw ensemble for ECC, includes many equal values, e.g., zeros for precipitation, reordering of quantiles of the postprocessed predictive distribution is not straightforward. Random reordering would lead to forecast scenarios with physically unrealistic jumps between spatially and temporally close locations. Resolving this issue may include methods by Scheuerer and Hamill (2018) to generate realistic forecast scenarios in the presence of ties or the modifications to ECC in order to smooth sharp jumps by Bellier et al. (2018).

The method of analogs already discussed in the "State-of-the-art statistical postprocessing methods" section is an approach which easily allows the introduction of spatial and temporal correlations in the output. For instance, comparing the entire fields or objects in the analog-search instead of independent point-by-point events and then associating the analog to an analysis provides a postprocessed solution possessing the spatial (and temporal) correlation of the analysis (e.g., Hamill and Whitaker 2006; Frediani et al. 2017).

Coping with model changes

Many of the methods presented in the preceding sections rely on the availability of large archives of past forecast and observation data, and it is usually assumed that the

error characteristics do not change substantially over time. However, this assumption is often threatened by inhomogeneous model and observation data brought about by changes in the observation systems and upgrades to NWP models. Techniques to homogenize observation data with the purpose of generating long observational time series exist, but the homogenization of NWP-model data is a more challenging problem.

To achieve ideal training datasets containing representations of many possible environmental conditions (including extremes) many past years of forecast data from ensemble prediction systems with a homogeneous design are required. The ideal solution would be to retrospectively generate past forecasts with the latest model version, restarting from previous model initial data, resulting in *reforecasts* (or *hindcasts*), all other features being equal. For example, reforecasts are provided by global prediction centers such as the U.S. National Weather Service or the European Centre for Medium-Range Weather Forecasts (Hamill et al. 2013). As discussed by Hamill (2018), this procedure is computationally very demanding and may impact many other aspects of the centers' activities. The added skill coming from additional training data for postprocessing methods thus needs to be weighed against the loss of computational resources for ensemble prediction system (EPS) model improvements, for example, by running the model at a higher resolution.

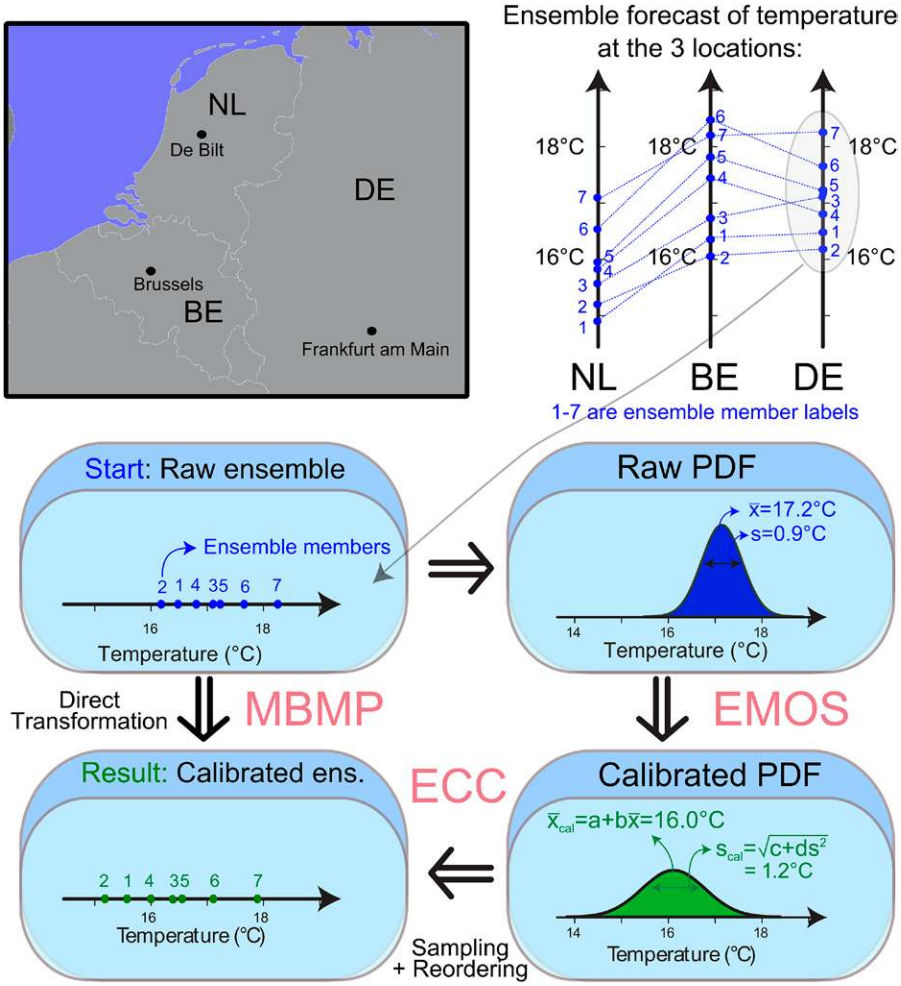


Fig. 2. Illustration of two methods to preserve correlations of high-dimensional corrected fields for each ensemble member, namely, the member-by-member postprocessing (MBMP) and EMOS-ECC. As an example, ensemble temperature forecasts are graphically shown for Frankfurt am Main [Germany (DE)], De Bilt [the Netherlands (NL)], and Brussels [Belgium (BE)]. (top right) The raw forecasts for each of the indicated ensemble members are correlated among the three locations and these correlations must also be present after calibration. Using a direct linear transformation, the MBMP method preserves them naturally, while the ECC is necessary to reorder samples from a calibrated probability density function produced using EMOS. From Schefzik (2017).

Many postprocessing schemes using time-adaptive training sets (e.g., sliding windows) have been developed to cope with that problem. They typically have to rely on past forecast cases from only the current and possibly a few past seasons (Wilson and Vallée 2002). Even though these methods adapt to new data being progressively added to the training data, eventually nullifying the need for a statistical model adaptation, temporary degradation of the postprocessing following model changes can often be observed (Lang et al. 2020), and may impact the overall performance of the statistical model. Therefore, the development of

methods to better incorporate model changes into postprocessing models has recently received substantial research interest. For instance, a new method based on the use of the trajectories of the linearized NWP model to possibly reduce the cost of homogenizing the past-forecasts database following a model change was recently proposed (Demaeyer and Vannitsem 2020). The effects of model changes will typically vary substantially according to the weather variable and context. For example, it may be essential for modeling rare events or small-scale phenomena to use a longer training period containing model changes rather than relying on a shorter period where reforecasts are available (Hess 2020).

The use of ML methods for postprocessing presents new opportunities to account for model changes during training (as well as suffering themselves from model changes). For example, binary indicators of changes of the NWP model versions could serve as additional predictors in the postprocessing model (Hess 2020). Further, information on the model change could be incorporated with linearized forecast models, possibly allowing ML algorithms to benefit from the trajectories of the linearized model (and the information of the model change encoded inside) as training data. Other methods like transfer learning should be tested in the present context; e.g., Ham et al. (2019).

Blending multiple forecasts

Conceptually, the blending of multiple forecast sources aims to create a single improved forecast (either by gaining continuity in transition or greater combined skill), based on the assumption that the inputs are samples from the same probability density function (PDF). In operational centers, the blending of multiple forecast sources has tended to be deterministic and focused on two forecasting concerns; first, optimizing the inclusion of a radar extrapolation nowcast (e.g., Golding 1998; Bailey et al. 2014) and second, incorporating multiple forecast sources (Woodcock and Engel 2005; Engel and Ebert 2012) to improve skill. An example of blending multiple model forecasts is provided in Fig. 3.

More recently the focus has included convection-permitting ensembles (Clark et al. 2016; Beck et al. 2016) and blending methods that are also capable of generating a single seamless forecast across time scales. Blending first requires a reprojection onto a common grid and either temporal interpolation or disaggregation to a common temporal frequency (Woodcock and Engel 2005). The inputs being blended should represent the same phenomena and that may require some form of calibration or downscaling (Howard and Clark 2007; Sheridan et al. 2010, 2018; Moseley 2011) to ensure equivalency between the inputs (e.g., Kober et al. 2012).

There are two fundamental ways of blending, either in the physical forecast space, or in probability space. In physical space, the simplest approach is to compute a lead time–dependent weighted average of multiple deterministic forecast sources (e.g., Haiden et al. 2011). For ensemble forecasts, all sources can be treated as equally likely members which simply increase the ensemble size (DelSole et al. 2013; Beck et al. 2016). A deterministic forecast can either be included as an additional member or an ensemble can be generated from the deterministic forecast, e.g., a nowcast extrapolation ensemble (Bowler et al. 2006; Atencia and Zawadzki 2014). Methods for generating an ensemble forecast from the observed radar truth include short-term EPS (STEPS) (Seed et al. 2013; Foresti et al. 2016), dynamically changing weight functions (Yu et al. 2015), and the use of an ensemble Kalman filter (Nerini et al. 2019). The physical realism of the resulting forecasts is key for providing inputs to applications, such as hydrological models (Berenguer et al. 2005; Heuvelink et al. 2020).

Probabilistic forecasts are derived by finding the probability of occurrence for a set of thresholds (e.g., $T > 0$°C, $T > 1$°C,...). Blending probabilistic forecasts produces a smooth blend, assuming similar properties for the input forecasts, by easily dealing with spatial mismatches between fields (Johnson and Wang 2012; Kober et al. 2012). The full ensemble distribution

is also retained, unlike for physical blending that tends to damp toward the climatological mean, while coherent scenarios in spatial structure can be regenerated. Blending probabilistic forecasts, as in Johnson and Wang (2012), Kober et al. (2012) and Bouttier and Marchal (2020), typically smooths the gridpoint forecasts using a neighborhood approach (Theis et al. 2005; Schwartz and Sobash 2017) to effectively generate more members and create smoother probability fields prior to calibrating using, for example, a reliability-based procedure (Zhu et al. 1996; Flowerdew 2014).

Optimizing a blend of forecast sources relies on the input forecasts having sufficient skill. Computationally expensive blending techniques are unlikely to be beneficial if the input forecasts are very poor. If the input forecasts are sufficiently skillful, then a simple blend can add more skill (Johnson and Swinbank 2009; Beck et al. 2016). Any blending approach involves weighting the inputs. Common ways are simple linear weighting approaches (Golding 1998; Woodcock and Engel 2005; Engel and Ebert 2012) and computing weights by optimizing a verification metric (Atencia et al. 2020; Bouttier and Marchal 2020; Schaumann et al. 2020). Tuning the weights by lead time, flow dependence (Atencia et al. 2010, 2020), spatial scale (Seed et al. 2013), convective cell identification (Feige et al. 2018; Posada et al. 2019), and regime classification (Kober et al. 2014) can add information, although for the optimal results this must be in conjunction with advanced calibration (Kober et al. 2014). Therefore, an approach in which multiple processing steps are tuned in unison (Bouttier and Marchal 2020) is appealing for optimizing the final result of the calibration and blending. When considering postprocessing multiple variables (Haiden et al. 2011), a strategy that also preserves the relationship between variables is desirable. Overall, the choice of physical or probability blending depends on the desired product. If coherent scenarios for driving hydrological models is the focus (Berenguer et al. 2005; Heuvelink et al. 2020), then a physical space blending is preferable, while blending in probability space is efficient, particularly, for operational systems processing a range of meteorological variables.

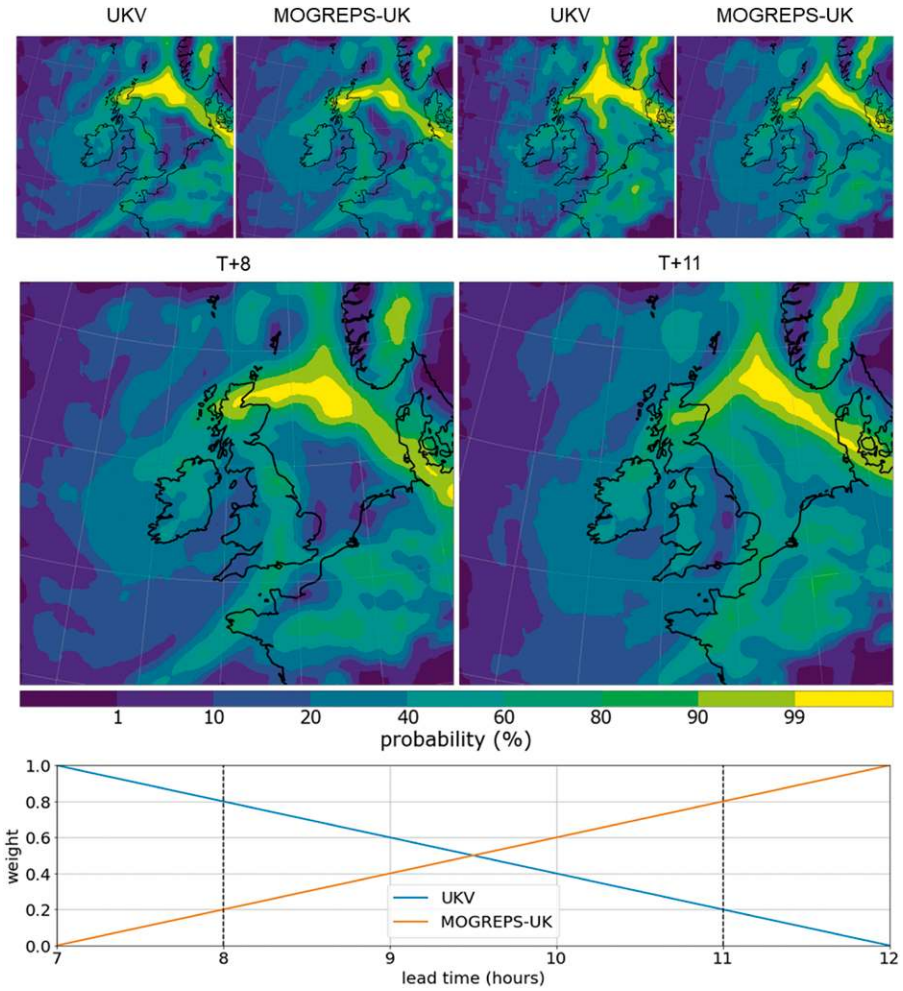


Fig. 3. Example of probabilistic blending for a domain centered on the United Kingdom. (top) The two NWP models [UKV (1.5 km) and MOGREPS-U.K. (2.2 km); Hagelin et al. 2017] used as inputs to the probabilistic blending on a 2 km grid at lead times of (top left) *T* + 8 and (top right) *T* + 11. (middle) The output from the probabilistic blending at the specified lead times constructed using a combination of UKV and MOGREPS-U.K. (bottom) The lead-time-dependent contributions to the blended output from the available forecast sources; the weighting used at each lead time is indicated by the dashed vertical lines.

To summarize, the key challenge of blending multiple forecast sources in physical space is the requirement to maintain realistic physical features within the blended output. This is a particular problem for more discrete fields, such as precipitation, where spatial mismatches between features within the input fields can lead to unphysical behavior in the blended output. Different forecast sources, especially if at very different resolutions, may not represent equivalent phenomena in the same way, which can then make the notion of blending questionable because they do not come from the same PDF. For blending in probability space, a key challenge remains the generation of physical scenarios from the probabilistic output that are coherent in space and time, and then a further difficulty may be retaining appropriate correlations between different variables. The ideal blending technique would be diagnostic agnostic, applicable across time scales and able to produce physically realistic forecasts, as required for driving downstream models. While various investigations have provided invaluable insights (Bowler et al. 2006; Johnson and Wang 2012; Kober et al. 2012; Seed et al. 2013), no single universally applicable and successful blending technique exists to date.

Challenges in operational implementation

For postprocessing, a key challenge is the availability of a training dataset within the operational environment with a truth of sufficient quality. Generation of larger training datasets is often prohibitive due to the data volume and timeliness requirements for operational running, and the presence of model upgrades (see the "Coping with model changes" section). For operational implementation, a smaller training dataset constructed for a rolling training period is therefore typically favored (Allen et al. 2019). Additionally, a choice is required as to whether calibrate gridded or site forecasts. NWP models output gridded forecasts, while high-quality observations as a source of truth are usually available at specific sites, only. This could pose problems when gridded calibrated outputs are sought for. The choice of calibrating gridded or site forecasts operationally is therefore hugely dependent on the reliable availability of a truth dataset of sufficient quality for calibration. Operational running mandates a compromise between the scientific ideals of the research community and the practicalities of managing a robust operational system.

The transfer from research to operations (R2O) requires considerable effort to configure a research framework into an operational software environment (see appendix for examples), ensuring proper run-time behavior, achieving acceptance by users, and maintaining the system in the long term. A gap between research and operations may be perceived, especially if the necessity for the R2O transfer is not sufficiently understood or acknowledged or the R2O process is not included in research projects. Other key challenges for operational implementation are (i) facilitating technical installation, (ii) ensuring data volumes and timeliness, (iii) producing outputs that are attractive for users, and (iv) maintaining the system in the long term.

Facilitating technical installation. The inclusion of statistical postprocessing into an operational forecast chain can add a great deal of complexity. Initial operational implementation requires several key elements listed here:

- Compliance with a technical production environment that is readily available at the NMS (e.g., coding language, data formats)
- Having suitable quality control and verification of the input and output data, especially in automated production to avoid unphysical customer products
- Ensuring robustness of the production pipelines
- Managing the structuring and archiving of large volumes of forecast and observational data for training

Further developments require communication with users, testing, and potentially time for adaptation by downstream users. To ensure a smooth flow of data for customers, separate development and operational production chains are necessary. During the development phase, performance would ideally be demonstrated in a trial using historic data, of sufficient length to probe multiple seasons and reliably measure performance for rarer events. The performance of the operational system needs to be similarly monitored to confirm the implementation provides similar performance to the trial, and to identify problems as they arise, record the quality of the data actually sent to customers, and build a further data archive that can be used to inform future development. New trials are needed when NWP models are changed.

Data volumes and timeliness. Operational weather forecasts (using both NWP and postprocessing) are constrained to run in a short period of time, since forecasts are useless if produced too late.

Traditionally, postprocessing is performed at station locations. This represents up to several thousand points, depending on country size and postprocessed element (many more for temperature than wind speed for example). But in many countries, there is a now a requirement for postprocessed fields on the model grid at kilometric scale, which may represent several million grid points. The timeliness problem is compounded by the introduction of more advanced machine learning techniques, since storing and loading into memory hundreds of regression trees for instance, instead of few coefficients is much more computationally expensive.

Taillardat and Mestre (2020) describe how stored elements may reach several thousand gigabytes of regression trees, when postprocessing several million points. These data can be handled by HPC through massively parallel computation capabilities, and extremely fast I/Os. Nevertheless, it takes a great deal of optimization to ensure proper use of HPC in real time.

It is however important to emphasize that classical regressions are also highly demanding when dealing with high resolution observations and forecasts, as the fitting is usually done locally at each grid point. ML can effectively cope with spatial heterogeneities by easily incorporating this information in the model structure (Scheuerer et al. 2020; Veldkamp et al. 2021). Other modifications of the ML formulation can also be done to reduce the storage needed.

Attractiveness for users. Attractiveness for users is essential for maintaining confidence in an operational capability. Users are looking for benefits from postprocessing that may vary. The exact definition of a "good" forecast and the most important technical characteristics depend on the user group and the communication channel used (e.g., visualization for forecasters, open data for external specialized users, smartphone app for the public). The main factors are as follows:

- Relevance and usability of the postprocessed output, i.e., choice of provided variables (including multivariate combinations), regions, output resolution (time/space), and lead times. In addition, there should be no inconsistencies, missing data, jumpiness, or unphysical output. There should also be a consideration of pragmatic requirements for which skill is not the only relevant factor and good visualization.
- Explanation of the output. This is usually realized by documentation, newsletters, FAQ tables, additional first level support, training courses, focus groups, etc.
- An interactive approach in which feedback from users includes responses from developers. A thoughtful agreement on the scope of the postprocessing and quality criteria should be made with the users. For instance, on the need for information at observed sites only, or

also at unobserved ones; a seamless appearance of the forecasts at different lead times; realistic variability in space and time; suitable verification scores for optimization; and coherence between deterministic and probabilistic forecasts.

Maintaining the system in the long term. Long-term maintenance also poses challenges. A key practical consideration is keeping the postprocessing system as portable as possible, allowing for ease of changes in operational infrastructure, for instance by using Docker. Another issue is maintaining the necessary in-house knowledge. Scientific and technical knowledge about the postprocessing system should be easily transferable from one person to another. This may be facilitated by good software development practices (i.e., readability, version management), and up-to-date technical and scientific documentation. Finally, adaptation to upcoming changes in input data and in the technical production environment at the NMS is vital, otherwise a postprocessing system can quickly become outdated. It means avoiding having too many different systems that all have overheads.

Future prospects on postprocessing

The operational implementation of statistical postprocessing techniques for multiple forecasts at very high spatial resolution and for time ranges from minutes to seasons is a challenge faced by all meteorological centers in the world. Despite improvements in NWP-model forecasts, the demand for local high-quality forecasts is increasing, and therefore, statistical postprocessing should be an integral part of the operational chain, and high priority is given to ongoing developments. This paper outlines the current advances and challenges faced by the meteorological community in this area.

A first key challenge is to effectively implement techniques that already provide, or have the potential to provide, important improvements to the forecasts (e.g., Hemri et al. 2014). There are currently considerable efforts in that direction at the meteorological services, usually in close collaboration with the academic research as illustrated in the second, third, and fourth sections. In view of the considerable increase of model resolution and processes involved, new tools should also be envisaged, such as the use of machine learning or blending approaches. Although these techniques have great potential for further improving forecasts, they should not be considered a panacea that will solve all the forecasting problems. No statistical method can go beyond the information that is contained within the input datasets, in particular if forecast trajectories are poor because of sensitivity to initial conditions. To really be able to benchmark the value of new methods, a common platform on which the different techniques can be compared on a set of appropriately chosen meteorological forecasts is highly desirable.

A second key challenge is the allocation of appropriate computational and staff resources. It may be simply impossible to implement some techniques. For instance, training and inference with comprehensive machine learning approaches can become a significant endeavor. Customized hardware to run NWP systems in use at most NMS and research centers may not be adequate for NWP postprocessing that is attempting to implement big data methods. Appropriate hardware and software should be allocated, such as the GPUs that provide opportunities for massively parallel implementation. Well-trained staff is also required for the research, development, and implementation.

A third challenge of NMS is to choose the appropriate techniques for their own or customers' purposes, as discussed in the "Challenges in operational implementation" section. There are a large number of possible approaches, and those chosen should be compatible with end-user purposes. As discussed in detail in the WMO publication on postprocessing (WMO 2021), simple approaches like bias corrections can be easily implemented even with very limited IT resources or manpower provided suitable observations are available, but

others are much more demanding like blending approaches, machine learning or analogs. The trade-off between complexity and requirements should be assessed beforehand.

A fourth challenge that was not touched upon in the present review is the use of new datasets provided for instance by crowdsourcing. These are proliferating rapidly and considerable efforts are now going into determining how to use them in NWP. Statistical postprocessing can of course benefit from these new sources of data for training the algorithms but also to allow more independent verification of the quality of the corrections. New datasets also introduce some new problems that need to be solved. For instance, when constructing a common calibration model over a large domain the observations may become highly variable in coverage. Furthermore, unstable availability may introduce unrealistic jumpiness for short lead times/nowcasting when updating calibrated forecasts.

A fifth challenge is to develop statistical postprocessing or blending approaches that could provide forecasts at any location, and not only at specific grid points or specific station sites. One option would be to include the location and variables derived from elevation and surface-type data as features in regression models in order to make them valid for whole domains. Another strategy would be to remove spatial heterogeneity in the training data before the statistical modeling and then retransform in the end (Dabernig et al. 2017). A third possibility is to apply spatial interpolation methods to well-calibrated forecasts at stations (Scheuerer and König 2014; Taillardat and Mestre 2020).

Several centers are regularly running reforecasts, but many questions arise about the number of ensemble members or number of past reforecasts needed for a specific method. The feedback from the postprocessing experts is an important challenge as it requires communication between the different meteorological centers, and to find compromise between the main needs and the forecasting capabilities.

Finally, there is a clear shift currently taking place from physical modeling approaches to data-driven approaches, due to the plethora of new datasets, new technologies and computer power resources, and data science techniques, that allow to improve the forecast quality by other means than improving the NWP models. The strategy and resource allocation of meteorological services should take this into account.

Acknowledgments. The constructive comments of the two reviewers were highly appreciated. Useful comments from Tim Hewson and Mark Liniger on an earlier version of this paper are also very much appreciated. The work of Jonathan Demaeyer and Stéphane Vannitsem is partly supported by the EUMETNET module "Post-processing" of the NWP cooperation program. Sebastian Lerch acknowledges support by the Deutsche Forschungsgemeinschaft through SFB/TRR 165 "Waves to Weather." Bert Van Schaeybroeck and Stéphane Vannitsem are partly supported by the project MEDSCOPE that has received funding from EU's H2020 Research and Innovation Program under Grant Agreement 690462.

Appendix: Available tools

The development of modern postprocessing methods is driven and facilitated by the availability of efficient, cross-platform open source software libraries such as ranger (Wright and Ziegler 2017) for random forests, or Keras (Chollet et al. 2015) and Tensorflow (Abadi et al. 2016) for neural networks. These modern tools complement available implementations of postprocessing schemes from research software packages such as the R packages crch (Messner et al. 2016), CSTools, or ensembleBMA (Fraley et al. 2011), all available on https://CRAN.R-project.org/. Operational routines at NMS are also available such as the Met Office's IMPROVER library (https://github.com/metoppv/improver; Evans et al. 2020) or the Finnish Meteorological Institute Himan tool (https://github.com/fmidev/himan). For a recent overview of computer software for postprocessing and usage examples, see Messner (2018).

References

Abadi, M., and Coauthors, 2016: Tensorflow: A system for large-scale machine learning. *Proc. USENIX 12th Symp. on Operating Systems Design and Implementation*, Savannah, GA, Advanced Computing Systems Association, 265–283, www.usenix.org/system/files/conference/osdi16/osdi16-abadi.pdf.

Alessandrini, S., L. Delle Monache, S. Sperati, and J. N. Nissen, 2015: A novel application of an analog ensemble for short-term wind power forecasting. *Renewable Energy*, **76**, 768–781, https://doi.org/10.1016/j.renene.2014.11.061.

——, ——, C. M. Rozoff, and W. E. Lewis, 2018: Probabilistic prediction of tropical cyclone intensity with an analog ensemble. *Mon. Wea. Rev.*, **146**, 1723–1744, https://doi.org/10.1175/MWR-D-17-0314.1.

Allen, S., C. A. T. Ferro, and F. Kwasniok, 2019: Regime-dependent statistical postprocessing of ensemble forecasts. *Quart. J. Roy. Meteor. Soc.*, **145**, 3535–3552, https://doi.org/10.1002/qj.3638.

Antun, V., F. Renna, C. Poon, B. Adcock, and A. C. Hansen, 2020: On instabilities of deep learning in image reconstruction and the potential costs of AI. *Proc. Natl. Acad. Sci. USA*, **117**, 30088–30095, https://doi.org/10.1073/pnas.1907377117.

Atencia, A., and I. Zawadzki, 2014: A comparison of two techniques for generating nowcasting ensembles. Part I: Lagrangian ensemble technique. *Mon. Wea. Rev.*, **142**, 4036–4052, https://doi.org/10.1175/MWR-D-13-00117.1.

——, and Coauthors, 2010: Improving QPF by blending techniques at the meteorological service of Catalonia. *Nat. Hazards Earth Syst. Sci.*, **10**, 1443–1455, https://doi.org/10.5194/nhess-10-1443-2010.

——, Y. Wang, A. Kann, and F. Meier, 2020: Localization and flow-dependency on blending techniques. *Meteor. Z.*, **29**, 231–246, https://doi.org/10.1127/metz/2019/0987.

Bailey, M. E., G. A. Isaac, I. Gultepe, I. Heckman, and J. Reid, 2014: Adaptive blending of model and observations for automated short-range forecasting: Examples from the Vancouver 2010 Olympic and Paralympic Winter Games. *Pure Appl. Geophys.*, **171**, 257–276, https://doi.org/10.1007/s00024-012-0553-x.

Baran, S., and S. Lerch, 2016: Mixture EMOS model for calibrating ensemble forecasts of wind speed. *Environmetrics*, **27**, 116–130, https://doi.org/10.1002/env.2380.

Bauer, P., A. Thorpe, and G. Brunet, 2015: The quiet revolution of numerical weather prediction. *Nature*, **525**, 47–55, https://doi.org/10.1038/nature14956.

Beck, J., F. Bouttier, L. Wiegand, C. Gebhardt, C. Eagle, and N. Roberts, 2016: Development and verification of two convection-allowing multi-model ensembles over western Europe. *Quart. J. Roy. Meteor. Soc.*, **142**, 2808–2826, https://doi.org/10.1002/qj.2870.

Bellier, J., I. Zin, and G. Bontron, 2018: Generating coherent ensemble forecasts after hydrological postprocessing: Adaptations of ECC-based methods. *Water Resour. Res.*, **54**, 5741–5762, https://doi.org/10.1029/2018WR022601.

Ben Bouallègue, Z., 2017: Statistical postprocessing of ensemble global radiation forecasts with penalized quantile regression. *Meteor. Z.*, **26**, 253–264, https://doi.org/10.1127/metz/2016/0748.

——, T. Heppelmann, S. E. Theis, and P. Pinson, 2016: Generation of scenarios from calibrated ensemble forecasts with a dual-ensemble copula-coupling approach. *Mon. Wea. Rev.*, **144**, 4737–4750, https://doi.org/10.1175/MWR-D-15-0403.1.

Berenguer, M., X. Corral, R. Sánchez-Diezma, and D. Sempere-Torres, 2005: Hydrological validation of a radar-based nowcasting technique. *J. Hydrometeor.*, **6**, 532–549, https://doi.org/10.1175/JHM433.1.

Bottou, L., and V. Vapnik, 1992: Local learning algorithms. *Neural Comput.*, **4**, 888–900, https://doi.org/10.1162/neco.1992.4.6.888.

Bouttier, F., and H. Marchal, 2020: Probabilistic thunderstorm forecasting by blending multiple ensembles. *Tellus*, **72A**, 1–19, https://doi.org/10.1080/16000870.2019.1696142.

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. *Quart. J. Roy. Meteor. Soc.*, **132**, 2127–2155, https://doi.org/10.1256/qj.04.100.

Breiman, L., 2001: Random forests. *Mach. Learn.*, **45**, 5–32, https://doi.org/10.1023/A:1010933404324.

Bremnes, J. B., 2004: Probabilistic forecasts of precipitation in terms of quantiles using NWP model output. *Mon. Wea. Rev.*, **132**, 338–347, https://doi.org/10.1175/1520-0493(2004)132<0338:PFOPIT>2.0.CO;2.

——, 2019: Constrained quantile regression splines for ensemble postprocessing. *Mon. Wea. Rev.*, **147**, 1769–1780, https://doi.org/10.1175/MWR-D-18-0420.1.

——, 2020: Ensemble postprocessing using quantile function regression based on neural networks and Bernstein polynomials. *Mon. Wea. Rev.*, **148**, 403–414, https://doi.org/10.1175/MWR-D-19-0227.1.

Bröcker, J., and L. A. Smith, 2008: From ensemble forecasts to predictive distribution functions. *Tellus*, **60A**, 663–678, https://doi.org/10.1111/j.1600-0870.2008.00333.x.

Buizza, R., 2019: Introduction to the special issue on "25 years of ensemble forecasting." *Quart. J. Roy. Meteor. Soc.*, **145**, 1–11, https://doi.org/10.1002/qj.3370.

——, and M. Leutbecher, 2015: The forecast skill horizon. *Quart. J. Roy. Meteor. Soc.*, **141**, 3366–3382, https://doi.org/10.1002/qj.2619.

Cannon, A. J., 2018: Non-crossing nonlinear regression quantiles by monotone composite quantile regression neural network, with application to rainfall extremes. *Stochastic Environ. Res. Risk Assess.*, **32**, 3207–3225, https://doi.org/10.1007/s00477-018-1573-6.

Cervone, G., L. Clemente-Harding, S. Allessandrini, and L. Delle-Monache, 2017: Short-term photovoltaic power forecasting using artificial neural networks and an analog ensemble. *Renewable Energy*, **108**, 274–286, https://doi.org/10.1016/j.renene.2017.02.052.

Chollet, F., and Coauthors, 2015: Keras: The Python deep learning library. Keras, https://keras.io.

Clark, M., S. Gangopadhyay, L. E. Hay, B. Rajagopalan, and R. L. Wilby, 2004: The Schaake shuffle: A method for reconstructing space–time variability in forecasted precipitation and temperature fields. *J. Hydrometeor.*, **5**, 243–262, https://doi.org/10.1175/1525-7541(2004)005<0243:TSSAMF>2.0.CO;2.

Clark, P., N. Roberts, H. Lean, S. P. Ballard, and C. Charlton-Perez, 2016: Convection-permitting models: A step-change in rainfall forecasting. *Meteor. Appl.*, **23**, 165–181, https://doi.org/10.1002/met.1538.

Cloke, H. L., and F. Pappenberger, 2009: Ensemble flood forecasting: A review. *J. Hydrol.*, **375**, 613–626, https://doi.org/10.1016/j.jhydrol.2009.06.005.

Collins, M., and M. R. Allen, 2002: Assessing the relative roles of initial and boundary conditions in interannual to decadal climate predictability. *J. Climate*, **15**, 3104–3109, https://doi.org/10.1175/1520-0442(2002)015<3104:ATRROI>2.0.CO;2.

Dabernig, M., G. J. Mayr, J. W. Messner, and A. Zeileis, 2017: Spatial ensemble post-processing with standardized anomalies. *Quart. J. Roy. Meteor. Soc.*, **143**, 909–916, https://doi.org/10.1002/qj.2975.

——, I. Schicker, A. Kann, Y. Wang, and M. N. Lang, 2020: Statistical post-processing with standardized anomalies based on a 1 km gridded analysis. *Meteor. Z.*, **29**, 265–275, https://doi.org/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. *Mon. Wea. Rev.*, **139**, 3554–3570, https://doi.org/10.1175/2011MWR3653.1.

——, T. Eckel, D. Rife, and B. Nagarajan, 2013: Probabilistic weather prediction with an analog ensemble. *Mon. Wea. Rev.*, **141**, 3498–3516, https://doi.org/10.1175/MWR-D-12-00281.1.

DelSole, T., X. Yang, and M. K. Tippett, 2013: Is unequal weighting significantly better than equal weighting for multi-model forecasting? *Quart. J. Roy. Meteor. Soc.*, **139**, 176–183, https://doi.org/10.1002/qj.1961.

Demaeyer, J., and S. Vannitsem, 2020: Correcting for model changes in statistical post-processing—An approach based on response theory. *Nonlinear Processes Geophys.*, **27**, 307–327, https://doi.org/10.5194/npg-27-307-2020.

Doblas-Reyes, F. J., R. Hagedorn, and T. N. Palmer, 2005: The rationale behind the success of multi-model ensembles in seasonal forecasting—II. Calibration and combination. *Tellus*, **57A**, 234–252, https://doi.org/10.3402/tellusa.v57i3.14658.

Engel, C., and E. E. Ebert, 2012: Gridded operational consensus forecasts of 2-m temperature over Australia. *Wea. Forecasting*, **27**, 301–322, https://doi.org/10.1175/WAF-D-11-00069.1.

Evans, G. R., and Coauthors, 2020: metoppv/IMPROVER: IMPROVER: A library of algorithms for meteorological post-processing, version 0.10.0. Zenodo, https://doi.org/10.5281/zenodo.3744431.

Feige, K., R. Posada, and U. Blahak, 2018: Developing a Concept to Visualize Object-based Weather Forecasting Ensembles. *Workshop on Visualisation in Environmental Sciences (EnvirVis)*, K. Rink, D. Zeckzer, R. Bujack, and S. Jänicke, Eds., The Eurographics Association, 19–25, http://dx.doi.org/10.2312/envirvis.20181133.

Flowerdew, J., 2014: Calibrating ensemble reliability whilst preserving spatial structure. *Tellus*, **66A**, 22662, https://doi.org/10.3402/tellusa.v66.22662.

Foresti, L., M. Reyniers, A. Seed, and L. Delobbe, 2016: Development and verification of a real-time stochastic precipitation nowcasting system for urban hydrology in Belgium. *Hydrol. Earth Syst. Sci.*, **20**, 505–527, https://doi.org/10.5194/hess-20-505-2016.

Fraley, C., A. Raftery, T. Gneiting, M. Sloughter, and V. Berrocal, 2011: Probabilistic weather forecasting in R. *R J.*, **3**, 55–63, https://doi.org/10.32614/RJ-2011-009.

Frediani, M. E. B., T. M. Hopson, J. P. Hacker, E. N. Anagnostou, L. Delle Monache, and F. Vandenberghe, 2017: Object-Based Analog Forecasts for Surface Wind Speed. *Mon. Wea. Rev.*, **145**, 5083–5102, https://doi.org/10.1175/MWR-D-17-0012.1.

Frogner, I. L., A. T. Singleton, M. Ø. Køltzow, and U. Andrae, 2019: Convection-permitting ensembles: Challenges related to their design and use. *Quart. J. Roy. Meteor. Soc.*, **145**, 90–106, https://doi.org/10.1002/qj.3525.

Gagne, D. J., A. McGovern, and J. Brotzge, 2009: Classification of convective areas using decision trees. *J. Atmos. Oceanic Technol.*, **26**, 1341–1353, https://doi.org/10.1175/2008JTECHA1205.1.

——, ——, S. E. Haupt, R. A. Sobash, J. K. Williams, and M. Xue, 2017: Storm-based probabilistic hail forecasting with machine learning applied to convection-allowing ensembles. *Wea. Forecasting*, **32**, 1819–1840, https://doi.org/10.1175/WAF-D-17-0010.1.

Gebetsberger, M., J. W. Messner, G. J. Mayr, and A. Zeileis, 2017: Fine-tuning non-homogeneous regression for probabilistic precipitation forecasts: Unanimous predictions, heavy tails, and link functions. *Mon. Wea. Rev.*, **145**, 4693–4708, https://doi.org/10.1175/MWR-D-16-0388.1.

Glahn, H. R., and D. A. Lowry, 1972: The use of model output statistics (MOS) in objective weather forecasting. *J. Appl. Meteor.*, **11**, 1203–1211, https://doi.org/10.1175/1520-0450(1972)011<1203:TUOMOS>2.0.CO;2.

Gneiting, T., and A. E. Raftery, 2007: Strictly proper scoring rules, prediction, and estimation. *J. Amer. Stat. Assoc.*, **102**, 359–378, https://doi.org/10.1198/016214506000001437.

——, ——, A. H. Westveld III, and T. Goldman, 2005: Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Mon. Wea. Rev.*, **133**, 1098–1118, https://doi.org/10.1175/MWR2904.1.

——, F. Balabdaoui, and A. E. Raftery, 2007: Probabilistic forecasts, calibration and sharpness. *J. Roy. Stat. Soc.*, **69B**, 243–268, https://doi.org/10.1111/j.1467-9868.2007.00587.x.

Golding, B. W., 1998: Nimrod: A system for generating automated very short range forecasts. *Meteor. Appl.*, **5**, 1–16, https://doi.org/10.1017/S1350482798000577.

Hagelin, S., J. Son, R. Swinbank, A. McCabe, N. Roberts, and W. Tennant, 2017: The Met Office convective-scale ensemble, MOGREPS-UK. *Quart. J. Roy. Meteor. Soc.*, **143**, 2846–2861, https://doi.org/10.1002/qj.3135.

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. *Wea. Forecasting*, **26**, 166–183, https://doi.org/10.1175/2010WAF2222451.1.

Ham, Y., J. Kim, and J. Luo, 2019: Deep learning for multi-year ENSO forecasts. *Nature*, **573**, 568–572, https://doi.org/10.1038/s41586-019-1559-7.

Hamill, T. M., 2018: Practical aspects of statistical postprocessing. *Statistical Post-processing of Ensemble Forecasts*, S. Vannitsem, D. S. Wilks, and J. Messner, Eds., Elsevier, 187–218.

——, and J. S. Whitaker, 2006: Probabilistic quantitative precipitation forecasts based on reforecast analogs: Theory and application. *Mon. Wea. Rev.*, **134**, 3209–3229, https://doi.org/10.1175/MWR3237.1.

——, R. Hagedorn, and J. S. Whitaker, 2008: Probabilistic forecast calibration using ECMWF and GFS ensemble reforecasts. Part II: Precipitation. *Mon. Wea. Rev.*, **136**, 2620–2632, https://doi.org/10.1175/2007MWR2411.1.

—, G. T. Bates, J. S. Whitaker, D. R. Murray, M. Fiorino, T. J. Galarneau Jr., Y. Zhu, and W. Lapenta, 2013: NOAA's second-generation global medium-range ensemble reforecast dataset. *Bull. Amer. Meteor. Soc.*, **94**, 1553–1565, https://doi.org/10.1175/BAMS-D-12-00014.1.

——, M. Scheuerer, and G. T. Bates, 2015: Analog probabilistic precipitation forecasts using GEFS reforecasts and climatology-calibrated precipitation analyses. *Mon. Wea. Rev.*, **143**, 3300–3309, https://doi.org/10.1175/MWR-D-15-0004.1.

Hemri, S., M. Scheuerer, F. Pappenberger, K. Bogner, and T. Haiden, 2014: Trends in the predictive performance of raw ensemble weather forecasts. *Geophys. Res. Lett.*, **41**, 9197–9205, https://doi.org/10.1002/2014GL062472.

Henzi, A., J. F. Ziegel, and T. Gneiting, 2019: Isotonic distributional regression. arXiv, https://arxiv.org/abs/1909.03725.

Hess, R., 2020: Statistical postprocessing of ensemble forecasts for severe weather at Deutscher Wetterdienst. *Nonlinear Processes Geophys.*, **27**, 473–487, https://doi.org/10.5194/npg-27-473-2020.

Heuvelink, D., M. Berenguer, C. C. Brauer, and R. Uijlenhoet, 2020: Hydrological application of radar rainfall nowcasting in the Netherlands. *Environ. Int.*, **136**, 105431, https://doi.org/10.1016/j.envint.2019.105431.

Hewson, T. D., and F. M. Pillosu, 2020: A new low-cost technique improves weather forecasts across the world. arXiv, https://arxiv.org/abs/2003.14397.

Howard, T., and P. Clark, 2007: Correction and downscaling of NWP wind speed forecasts. *Meteor. Appl.*, **14**, 105–116, https://doi.org/10.1002/met.12.

Johnson, A., and X. Wang, 2012: Verification and calibration of neighborhood and object-based probabilistic precipitation forecasts from a multimodel convection-allowing ensemble. *Mon. Wea. Rev.*, **140**, 3054–3077, https://doi.org/10.1175/MWR-D-11-00356.1.

Johnson, C., and N. Bowler, 2009: On the reliability and calibration of ensemble forecasts. *Mon. Wea. Rev.*, **137**, 1717–1720, https://doi.org/10.1175/2009MWR2715.1.

——, and R. Swinbank, 2009: Medium-range multimodel ensemble combination and calibration. *Quart. J. Roy. Meteor. Soc.*, **135**, 777–794, https://doi.org/10.1002/qj.383.

Junk, C., L. Delle Monache, and S. Alessandrini, 2015: Analog-based ensemble model output statistics. *Mon. Wea. Rev.*, **143**, 2909–2917, https://doi.org/10.1175/MWR-D-15-0095.1.

Kleiber, W., A. E. Raftery, J. Baars, T. Gneiting, C. Mass, and E. P. Grimit, 2011: Locally calibrated probabilistic temperature forecasting using geostatistical model averaging and local Bayesian model averaging. *Mon. Wea. Rev.*, **139**, 2630–2649, https://doi.org/10.1175/2010MWR3511.1.

Klein, W. H., B. M. Lewis, and I. Enger, 1959: Objective prediction of five-day mean temperatures during winter. *J. Meteor.*, **16**, 672–682, https://doi.org/10.1175/1520-0469(1959)016<0672:OPOFDM>2.0.CO;2.

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. *Quart. J. Roy. Meteor. Soc.*, **138**, 755–768, https://doi.org/10.1002/qj.939.

——, ——, and ——, 2014: Aspects of short-term probabilistic blending in different weather regimes. *Quart. J. Roy. Meteor. Soc.*, **140**, 1179–1188, https://doi.org/10.1002/qj.2220.

Lang, M. N., G. J. Mayr, R. Stauffer, and A. Zeileis, 2019: Bivariate Gaussian models for wind vectors in a distributional regression framework. *Adv. Stat. Climatol. Meteor. Oceanogr.*, **5**, 115–132, https://doi.org/10.5194/ascmo-5-115-2019.

——, S. Lerch, G. J. Mayr, T. Simon, R. Stauffer, and A. Zeileis, 2020: Remember the past: A comparison of time-adaptive training schemes for non-homogeneous

regression. *Nonlinear Processes Geophys.*, **27**, 23–34, https://doi.org/10.5194/npg-27-23-2020.

Lemcke, C., and S. Kruizinga, 1988: Model output statistics forecasts: Three years of operational experience in the Netherlands. *Mon. Wea. Rev.*, **116**, 1077–1090, https://doi.org/10.1175/1520-0493(1988)116<1077:MOSFTY>2.0.CO;2.

Lerch, S., and S. Baran, 2017: Similarity-based semilocal estimation of post-processing models. *J. Roy. Stat. Soc.*, **66C**, 21–51, https://doi.org/10.1111/rssc.12153.

Leutbecher, M., and T. N. Palmer, 2008: Ensemble forecasting. *J. Comput. Phys.*, **227**, 3515–3539, https://doi.org/10.1016/j.jcp.2007.02.014.

Loader, C., 1999: *Local Regression and Likelihood*. Springer Verlag, 290 pp.

Lorenz, E. N., 1982: Atmospheric predictability experiments with a large numerical model. *Tellus*, **34**, 505–513, https://doi.org/10.3402/tellusa.v34i6.10836.

Lynch, P., 2008: The origins of computer weather prediction and climate modeling. *J. Comput. Phys.*, **227**, 3431–3444, https://doi.org/10.1016/j.jcp.2007.02.034.

Marzban, C., 2003: Neural networks for postprocessing model output: ARPS. *Mon. Wea. Rev.*, **131**, 1103–1111, https://doi.org/10.1175/1520-0493(2003)131<1103:NNFPMO>2.0.CO;2.

McGovern, A., K. L. Elmore, D. J. Gagne, S. E. Haupt, C. D. Karstens, R. Lagerquist, T. Smith, and J. K. Williams, 2017: Using artificial intelligence to improve real-time decision-making for high-impact weather. *Bull. Amer. Meteor. Soc.*, **98**, 2073–2090, https://doi.org/10.1175/BAMS-D-16-0123.1.

——, R. Lagerquist, D. J. Gagne, G. E. Jergensen, K. L. Elmore, C. R. Homeyer, and T. Smith, 2019: Making the black box more transparent: Understanding the physical implications of machine learning. *Bull. Amer. Meteor. Soc.*, **100**, 2175–2199, https://doi.org/10.1175/BAMS-D-18-0195.1.

Meinshausen, N., 2006: Quantile regression forests. *J. Mach. Learn. Res.*, **7**, 983–999, https://jmlr.org/papers/v7/meinshausen06a.html.

Messner, J. W., 2018: Ensemble postprocessing with R. *Statistical Postprocessing of Ensemble Forecasts*, S. Vannitsem, D. S. Wilks, and J. Messner, Eds., Elsevier, 291–329.

Messner, J. W., G. J. Mayr, and A. Zeileis, 2016: Heteroscedastic censored and truncated regression with crch. *R J.*, **8**, 173–181, https://doi.org/10.32614/RJ-2016-012.

——, ——, and ——, 2017: Nonhomogeneous boosting for predictor selection in ensemble postprocessing. *Mon. Wea. Rev.*, **145**, 137–147, https://doi.org/10.1175/MWR-D-16-0088.1.

Möller, A., L. Spazzini, D. Kraus, T. Nagler, and C. Czado, 2018: Vine copula based post-processing of ensemble forecasts for temperature. arXiv, https://arxiv.org/abs/1811.02255.

Molnar, C., 2019: Interpretable machine learning: A guide for making black box models explainable. GitHub, https://christophm.github.io/interpretable-ml-book.

Molteni, F., R. Buizza, T. N. Palmer, and T. Petroliagis, 1996: The ECMWF Ensemble Prediction System: Methodology and validation. *Quart. J. Roy. Meteor. Soc.*, **122**, 73–119, https://doi.org/10.1002/qj.49712252905.

Moseley, S., 2011: From observations to forecasts—Part 12: Getting the most out of model data. *Weather*, **66**, 272–276, https://doi.org/10.1002/wea.844.

Nagarajan, B., L. Delle Monache, J. P. Hacker, D. L. Rife, K. Searight, J. C. Knieveland, and T. Nipen, 2015: An evaluation of analog-based postprocessing methods across several variables and forecast models. *Wea. Forecasting*, **30**, 1623–1643, https://doi.org/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. *Mon. Wea. Rev.*, **147**, 987–1006, https://doi.org/10.1175/MWR-D-18-0258.1.

Nicolis, C., 2007: Dynamics of model error: The role of the boundary conditions. *J. Atmos. Sci.*, **64**, 204–215, https://doi.org/10.1175/JAS3806.1.

——, R. Perdigao, and S. Vannitsem, 2009: Dynamics of prediction errors under the combined effect of initial condition and model errors. *J. Atmos. Sci.*, **66**, 766–778, https://doi.org/10.1175/2008JAS2781.1.

Odak Plenković, I., L. Delle Monache, K. Horvath, and M. Hrastinski, 2018: Deterministic wind speed predictions with analog-based methods over complex topography. *J. Appl. Meteor. Climatol.*, **57**, 2047–2070, https://doi.org/10.1175/JAMC-D-17-0151.1.

——, I. Schicker, M. Dabernig, K. Horvath, and E. Keresturi, 2020: Analog-based post-processing of the ALADIN-LAEF ensemble predictions in complex terrain. *Quart. J. Roy. Meteor. Soc.*, **146**, 1842–1860, https://doi.org/10.1002/qj.3769.

Pillosu, F., and T. Hewson, 2017: New point-rainfall forecasts for flash flood prediction. *ECMWF Newsletter*, No. 153, ECMWF, Reading, United Kingdom, 2–3, www.ecmwf.int/en/newsletter/153/news/new-point-rainfall-forecasts-flash-flood-prediction.

Pinson, P., and R. Girard, 2012: Evaluating the quality of scenarios of short-term wind power generation. *Appl. Energy*, **96**, 12–20, https://doi.org/10.1016/j.apenergy.2011.11.004.

——, and J. W. Messner, 2018: Application of postprocessing for renewable energy. *Statistical Postprocessing of Ensemble Forecasts*, S. Vannitsem, D. S. Wilks, and J. W. Messner, Eds., Elsevier, 241–266.

Posada, R., R. Feger, M. Schultze, K. Wapler, and M. Werner, 2019: Combination of object-based probabilistic Nowcasting and NWP-Ensemble. *10th European Conf. on Severe Storms (ECSS)*, Krakow, Poland, European Severe Storms Laboratory.

Raftery, A. E., T. Gneiting, F. Balabdaoui, and M. Polakowski, 2005: Using Bayesian model averaging to calibrate forecast ensembles. *Mon. Wea. Rev.*, **133**, 1155–1174, https://doi.org/10.1175/MWR2906.1.

Rasp, S., and S. Lerch, 2018: Neural networks for postprocessing ensemble weather forecasts. *Mon. Wea. Rev.*, **146**, 3885–3900, https://doi.org/10.1175/MWR-D-18-0187.1.

Richardson, D. S., 2000: Skill and economic value of the ECMWF Ensemble Prediction System. *Quart. J. Roy. Meteor. Soc.*, **126**, 649–667, https://doi.org/10.1002/qj.49712656313.

Rigby, R. A., and D. M. Stasinopoulos, 2005: Generalized additive models for location, scale and shape. *J. Roy. Stat. Soc.*, **54C**, 507–554, https://doi.org/10.1111/j.1467-9876.2005.00510.x.

Roulston, M. S., and L. A. Smith, 2003: Combining dynamical and statistical ensembles. *Tellus*, **55A**, 16–30, https://doi.org/10.3402/tellusa.v55i1.12082.

Schaumann, P., M. De Langlard, R. Hess, P. James, and V. Schmidt, 2020: A calibrated combination of probabilistic precipitation forecasts to achieve a seamless transition from nowcasting to very-short-range forecasting. *Wea. Forecasting*, **35**, 773–791, https://doi.org/10.1175/WAF-D-19-0181.1.

Schefzik, R., 2016: A similarity-based implementation of the Schaake shuffle. *Mon. Wea. Rev.*, **144**, 1909–1921, https://doi.org/10.1175/MWR-D-15-0227.1.

——, 2017: Ensemble calibration with preserved correlations: Unifying and comparing ensemble copula coupling and member-by-member postprocessing. *Quart. J. Roy. Meteor. Soc.*, **143**, 999–1008, https://doi.org/10.1002/qj.2984.

——, and A. Möller, 2018: Ensemble postprocessing methods incorporating dependence structures. *Statistical Postprocessing of Ensemble Forecasts*, S. Vannitsem, D. Wilks, and J. Messner, Eds., Elsevier, 91–125.

——, T. L. Thorarinsdottir, T. Gneiting: 2013: Uncertainty Quantification in Complex Simulation Models Using Ensemble Copula Coupling. *Statist. Sci.*, **28**, 616–640, https://doi.org/10.1214/13-STS443.

Scheuerer, M., and G. König, 2014: Gridded, locally calibrated, probabilistic temperature forecasts based on ensemble model output statistics. *Quart. J. Roy. Meteor. Soc.*, **140**, 2582–2590, https://doi.org/10.1002/qj.2323.

——, and T. M. Hamill, 2018: Generating calibrated ensembles of physically realistic, high-resolution precipitation forecast fields based on GEFS model output. *J. Hydrometeor.*, **19**, 1651–1670, https://doi.org/10.1175/JHM-D-18-0067.1.

——, and ——, 2019: Probabilistic forecasting of snowfall amounts using a hybrid between a parametric and an analog approach. *Mon. Wea. Rev.*, **147**, 1047–1064, https://doi.org/10.1175/MWR-D-18-0273.1.

——, ——, B. Whitin, M. He, and A. Henkel, 2017: A method for preferential selection of dates in the Schaake shuffle approach to constructing spatiotemporal forecast fields of temperature and precipitation. *Water Resour. Res.*, **53**, 3029–3046, https://doi.org/10.1002/2016WR020133.

——, M. B. Switanek, R. P. Worsnop, and T. M. Hamill, 2020: Using artificial neural networks for generating probabilistic subseasonal precipitation forecasts over California. *Mon. Wea. Rev.*, **148**, 3489–3506, https://doi.org/10.1175/MWR-D-20-0096.1.

Schlosser, L., T. Hothorn, R. Stauffer, and A. Zeileis, 2019: Distributional regression forests for probabilistic precipitation forecasting in complex terrain. *Ann. Appl. Stat.*, **13**, 1564–1589, https://doi.org/10.1214/19-AOAS1247.

Schwartz, C. S., and R. A. Sobash, 2017: Generating probabilistic forecasts from convection-allowing ensembles using neighborhood approaches: A review and recommendations. *Mon. Wea. Rev.*, **145**, 3397–3418, https://doi.org/10.1175/MWR-D-16-0400.1.

——, G. S. Romine, R. A. Sobash, K. R. Fossell, and M. L. Weisman, 2019: NCAR's real-time convection-allowing ensemble project. *Bull. Amer. Meteor. Soc.*, **100**, 321–343, https://doi.org/10.1175/BAMS-D-17-0297.1.

Seed, A. W., C. E. Pierce, and K. Norman, 2013: Formulation and evaluation of a scale decomposition-based stochastic precipitation nowcast scheme. *Water Resour. Res.*, **49**, 6624–6641, https://doi.org/10.1002/wrcr.20536.

Sheridan, P., S. Smith, A. Brown, and S. Vosper, 2010: A simple height-based correction for temperature downscaling in complex terrain. *Meteor. Appl.*, **17**, 329–339, https://doi.org/10.1002/met.177.

——, S. Vosper, and S. Smith, 2018: A physically based algorithm for downscaling temperature in complex terrain. *J. Appl. Meteor. Climatol.*, **57**, 1907–1929, https://doi.org/10.1175/JAMC-D-17-0140.1.

Simon, T., G. J. Mayr, N. Umlauf, and A. Zeileis, 2019: NWP-based lightning prediction using flexible count data regression. *Adv. Stat. Climatol. Meteor. Oceanogr.*, **5**, 1–16, https://doi.org/10.5194/ascmo-5-1-2019.

Sperati, S., S. Alessandrini, and L. Delle Monache, 2017: Gridded probabilistic weather forecasts with an analog ensemble. *Quart. J. Roy. Meteor. Soc.*, **143**, 2874–2885, https://doi.org/10.1002/qj.3137.

Taillardat, M., and O. Mestre, 2020: From research to applications—Examples of operational ensemble post-processing in France using machine learning. *Nonlinear Processes Geophys.*, **27**, 329–347, https://doi.org/10.5194/npg-27-329-2020.

——, ——, M. Zamo, and P. Naveau, 2016: Calibrated ensemble forecasts using quantile regression forests and ensemble model output statistics. *Mon. Wea. Rev.*, **144**, 2375–2393, https://doi.org/10.1175/MWR-D-15-0260.1.

——, A. L. Fougères, P. Naveau, and O. Mestre, 2019: Forest-based and semiparametric methods for the postprocessing of rainfall ensemble forecasting. *Wea. Forecasting*, **34**, 617–634, https://doi.org/10.1175/WAF-D-18-0149.1.

Theis, S. E., A. Hense, and U. Damrath, 2005: Probabilistic precipitation forecasts from a deterministic model: A pragmatic approach. *Meteor. Appl.*, **12**, 257, https://doi.org/10.1017/S1350482705001763.

Thorarinsdottir, T. L., and N. Schuhen, 2018: Verification: Assessment of calibration and accuracy. *Statistical Postprocessing of Ensemble Forecasts*, S. Vannitsem, D. S. Wilks, and J. Messner, Eds., Elsevier, 155–186.

Toth, Z., and E. Kalnay, 1993: Ensemble forecasting at NMC: The generation of perturbations. *Bull. Amer. Meteor. Soc.*, **74**, 2317–2330, https://doi.org/10.1175/1520-0477(1993)074<2317:EFANTG>2.0.CO;2.

Vannitsem, S., 2017: Predictability of large-scale atmospheric motions: Lyapunov exponents and error dynamics. *Chaos*, **27**, 032101, https://doi.org/10.1063/1.4979042.

——, D. S. Wilks, and J. W. Messner, 2018: *Statistical Postprocessing of Ensemble Forecasts*. Elsevier, 347 pp.

Van Schaeybroeck, B., and S. Vannitsem, 2011: Post-processing through linear regression. *Nonlinear Processes Geophys.*, **18**, 147–160, https://doi.org/10.5194/npg-18-147-2011.

——, and ——, 2015: Ensemble post-processing using member-by-member approaches: Theoretical aspects. *Quart. J. Roy. Meteor. Soc.*, **141**, 807–818, https://doi.org/10.1002/qj.2397.

van Straaten, C., K. Whan, and M. Schmeits, 2018: Statistical postprocessing and multivariate structuring of high-resolution ensemble precipitation forecasts. *J. Hydrometeor.*, **19**, 1815–1833, https://doi.org/10.1175/JHM-D-18-0105.1.

——, ——, D. Coumou, B. van den Hurk, and M. Schmeits, 2020: The influence of aggregation and statistical post-processing on the subseasonal predictability of European temperatures. *Quart. J. Roy. Meteor. Soc.*, **146**, 2654–2670, https://doi.org/10.1002/qj.3810.

Veldkamp, S., K. Whan, S. Dirksen, and M. Schmeits, 2021: Statistical postprocessing of wind speed forecasts using convolutional neural networks. *Mon. Wea. Rev.*, https://doi.org/10.1175/MWR-D-20-0219.1, in press.

Velthoen, J., J. Cai, G. Jongbloed, and M. Schmeits, 2019: Improving precipitation forecasts using extreme quantile regression. *Extremes*, **22**, 599–622, https://doi.org/10.1007/s10687-019-00355-1.

Wahl, S., 2015: Uncertainty in mesoscale numerical weather prediction: Probabilistic forecasting of precipitation. Ph.D., University of Bonn, 120 pp., http://hss.ulb.uni-bonn.de/2015/4190/4190.htm.

Whan, K., and M. Schmeits, 2018: Comparing area probability forecasts of (extreme) local precipitation using parametric and machine learning statistical postprocessing methods. *Mon. Wea. Rev.*, **146**, 3651–3673, https://doi.org/10.1175/MWR-D-17-0290.1.

Wilks, D. S., 2011: *Statistical Methods in the Atmospheric Sciences*. 3rd ed. Elsevier, 676 pp.

Wilson, L. J., and M. Vallée, 2002: The Canadian Updateable Model Output Statistics (UMOS) system: Design and development tests. *Wea. Forecasting*, **17**, 206–222, https://doi.org/10.1175/1520-0434(2002)017<0206:TCUMOS>2.0.CO;2.

WMO, 2021: Guidelines for Ensemble Prediction System Products and Post-Processing (TT-EPSPP). WMO/TD-1254, 37 pp., in press.

Woodcock, F., and C. Engel, 2005: Operational consensus forecasts. *Wea. Forecasting*, **20**, 101–111, https://doi.org/10.1175/WAF-831.1.

Wright, M., and A. Ziegler, 2017: ranger: A fast implementation of random forests for high dimensional data in C++ and R. *J. Stat. Software*, **77**, 1–17, https://doi.org/10.18637/jss.v077.i01.

Yoden, S., 2007: Atmospheric Predictability. *J. Meteor. Soc. Japan. Ser. II*, **85B**, 77–102, doi.org/10.2151/jmsj.85B.77.

Yu, W., E. Nakakita, S. Kim, and K. Yamaguchi, 2015: Improvement of rainfall and flood forecasts by blending ensemble NWP rainfall with radar prediction considering orographic rainfall. *J. Hydrol.*, **531**, 494–507, https://doi.org/10.1016/j.jhydrol.2015.04.055.

Zamo, M., 2016: Statistical post-processing of deterministic and ensemble wind speed forecasts on a grid. Ph.D., Université Paris-Saclay, 166 pp., www.theses.fr/2016SACLA029.

Zhu, Y., G. Iyengar, Z. Toth, S. M. Tracton, and T. Marchok, 1996: Objective evaluation of the NCEP Global Ensemble Forecasting System. *15th Conf. on Weather Analysis and Forecasting*, Norfolk, VA, Amer. Meteor. Soc., J79–J82.