Chapter 7

Practical Aspects of Statistical Postprocessing

Thomas M. Hamill
NOAA Earth System Research Lab, Physical Sciences Division
tom.hamill@noaa.gov
Draft, 7 March 2017

ABSTRACT

Many readers of this text may be developing improved post-processing algorithms with the desire to have them used regularly, such as for the daily adjustment of real-time weather guidance produced by an operational weather prediction facility. This chapter discusses some of the practical aspects involved with the development and technology transition of advanced post-processing algorithms. Topics will include challenges involved with the preparation of high-quality training data sets and possible compromises one may wish to consider in an environment where the training data is limited. The chapter will also provide a case study and suggest some changes that the post-processing community can institute to more rapidly move advanced methodologies from research into regular operational use.

Keywords: reanalyses, reforecasts, data assimilation, bias-variance tradeoff, quantile mapping, ensemble dressing, supplemental locations.

7.1. Introduction.

Those involved with statistical model development commonly spend much of their time dealing with the practical aspects behind testing a research hypothesis. What data should be used? Is the input data of consistent quality, or must the researcher perform quality control? Is the training data so limited that no existing method produces acceptable quality guidance? Is it so voluminous as to be challenging to store and disseminate or to speedily train a model? Does the training data change in its statistical characteristics over time? How do I quickly obtain code for existing methods to use as standards of comparison? A researcher may wish to focus on the scientific aspects of the problem but find they cannot do so until due diligence has been paid to these other issues. Such issues will not go away, but it is possible to anticipate and surmount common obstacles, individually and as a community.

Figure 7.1 illustrates a typical weather prediction system with its components and data stores, illustrating the dependency of statistical postprocessing on previously produced data. These components commonly include a data assimilation system (Daley 1991, Kalnay 2003) that statistically adjusts prior numerical forecasts to newly available observations. Its purpose is to generate accurate and dynamically balanced gridded analyses of the state of the

environment suitable for the initialization of a prediction system. The forecast model (or more commonly now, an ensemble prediction system; see chapter 2) approximates the laws governing the evolution of the environmental state (Durran 2010, Warner 2011) and simulates the evolution from the initial states. The statistical post-processing algorithm is commonly trained using archives of forecast, observation, and/or analysis data.

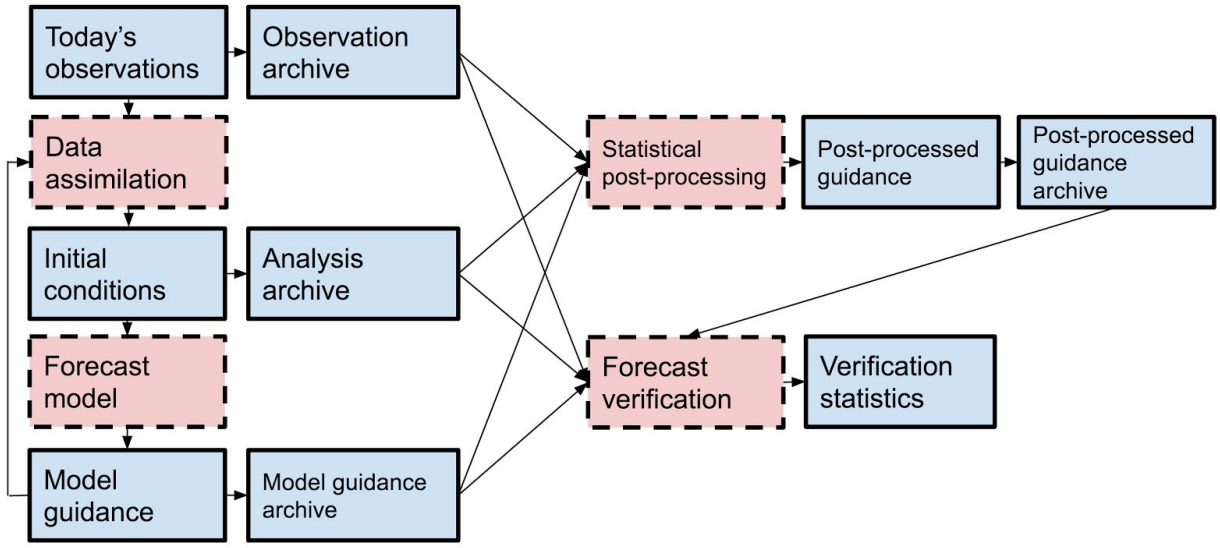

Figure 7.1: Diagram of many of the typical components and data stores of an end-to-end weather prediction system, and the propagation of data through the system. Boxes with solid borders are data stores, and boxes with dashed borders are components of the prediction system.

This diagram simplies the actual data flow. For example, statistical postprocessing often has two distinct phases, the training of a model and the application of that model to adjust today's real-time guidance. For some variables such as precipitation, the analyses used in the statistical training may (Lespinas et al. 2015) or may not (Zhang et al. 2016) utilize prior model forecast guidance, as suggested in the diagram. Further, the post-processed guidance is not necessarily the end of the product chain; it may also provide inputs to other prediction systems. For example, a hydrologic prediction system intended to produce streamflow forecasts may ingest post-processed meteorological guidance, synthesize it with observations of the land and snow state, and then generate ensembles of hydrologic predictions which in turn may require their own statistical postprocessing (Schaake et al. 2007).

Because of these data dependencies, the quality of the post-processed guidance depends on more than just the sophistication of the statistical algorithm. Suppose a statistical postprocessing algorithm is trained against analysis data, regarding these as proxies for the true state. The ultimate accuracy of the post-processed guidance thus depends upon the accuracy, bias, and temporal consistency of these analyses. Further, the post-processing algorithm is statistically modeling the discrepancies between prior forecasts and the verification data. What

should be done if the characteristics of the forecast discrepancies change in time due to something other than weather variability? Perhaps the forecast model has different bias characteristics in the warm season relative to the cool season, or El Niño vs La Niña conditions, or perhaps the forecast model was upgraded to a new version during the training period, and the old and new versions have different error characteristics. Understanding these issues and addressing them may be essential to providing the high-quality post-processed guidance desired by forecast users.

This chapter now delves more deeply into these issues and some possible ways to ameliorate them. Section 7.2 provides an example of how the classical and thorny "bias-variance tradeoff" manifests itself in statistical postprocessing; this tradeoff underlies the discussion of many of the algorithmic and data choices that follow. Section 7.3 then returns to discuss challenges with the training data, both forecast and observed/analyzed data. Section 7.4 discusses future directions to mitigate these challenges. Section 7.5 then provides a case study, discussing the tradeoffs that were made in developing a product of common interest, the probability of precipitation from multi-model ensemble guidance. Finally, in section 7.6 we turn to a different problem: how do we accelerate progress in statistical post-processing as a community? Different investigators commonly develop methods in isolation from each other, which may make testing a hypothesis (is the proposed method better than other recently developed methods?) quite difficult. There is a way forward, providing we are willing to participate in the co-development of a community infrastructure and test data sets.

7.2: The bias-variance tradeoff.

Reader are referred to applied statistics texts such as Hastie and Tibshirani (1990, Fig. 2.2) or Hastie et al. (2001, section 2.9) for more discussion on this subject. The bias-variance tradeoff is intimately related to a statistical concept called "overfitting." For example, this is discussed in Wilks (2011, section 7.4). Wikipedia (2016) describes the bias-variance tradeoff this way:

> *"The bias–variance tradeoff is a central problem in supervised learning[1]. Ideally, one wants to choose a model that both accurately captures the regularities in its training data, but also generalizes well to unseen data. Unfortunately, it is typically impossible to do both simultaneously. High-variance learning methods may be able to represent their training set well, but are at risk of overfitting to noisy or unrepresentative training data. In contrast, algorithms with high bias typically produce simpler models that don't tend to overfit, but may underfit their training data, failing to capture important regularities."*

Let's construct a simple, synthetic observation and forecast training data set to illustrate the problem that occurs with a commonly applied statistical post-processing algorithm, a "decaying-average bias correction" (Cui et al. 2012). Today's forecast bias is estimated as a

[1] Supervised learning is the machine learning task of inferring a function from labeled training data.

linear combination of the most recent forecast minus observation and a previous bias estimate. This simple post-processing approach is appealing for its minimal data storage requirements. Our theoretical construct is as follows. The true state of a univariate system at date/time *t*, y_t^{true} is sought. In this synthetic construct, the true state, unknown for purposes of model training, is always exactly zero. What is available is a time series of forecasts, all for the same lead time (say, perhaps a 3-day ahead forecast) from dates/times *t0* to *tf*, **x** $= [x_{t0}, \ldots, x_{tf}]$. Past observations **y** = $[y_{t0}, \ldots, y_{tf-1}]$ are also available. The observations are generated from the truth plus random noise: $y_t^o = y_t^{true} + e_t^o, \quad e_t^o \sim N(0, \frac{1}{9}),$ that is, the observations at time *t* are normally distributed with zero mean (the true state) and random error with a variance of 1/9. Forecast-error characteristics, unknown to the data analyst but known to us here, are constructed with random, seasonally dependent, and serially correlated systematic errors. The true seasonally dependent bias is $B_t = cos(2\pi J(t)/365)$, where *J(t)* is the Julian day of the year minus one; that is, the bias varies over the year from 1 to -1 in a cosine-shaped function, too warm at the beginning and end of the calendar year and too cold in the middle. The forecast's daily random error $e_t^f \sim N(0,1)$, i.e, the innovation variance (Wilks 2011, section 9.3.1) is here nine times larger than the observation variance. Finally, the time series of synthetic forecasts are simulated with a first-order autoregressive model (ibid) : $x_t - B_t = k(x_{t-1} - B_{t-1}) + e_t^f$, where here *k* = 0.5.

The decaying-average bias correction assumes that estimated forecast bias for day *t*, $\widehat{B}_t$, can be estimated as a linear combination of the previous day's bias estimate and the most recent deviation of the forecast from the observation:

$$\widehat{B}_t = (1-\alpha)\widehat{B}_{t-1} + \alpha(x_{t-1} - y_{t-1}^o). \qquad (7.1)$$

Here α is a user-defined parameter that indicates how much weight to apply to the most recent deviation of the observation from the forecast. When α is small, the bias tends toward being estimated as a long-term mean of the difference between forecasts and observations. When α is large, the most recent data is weighted heavily, and estimated bias may vary a lot from one day to the next.

Figure 7.2 illustrates 100 independent Monte-Carlo simulations of the estimated bias started from different initial random numbers and using different random observation errors; data is shown only after 60 days of spinup. The four panels show the simulations for four increasing values of α. Each simulation's estimated bias is shown with a light gray line. The mean of these bias estimates is shown with the dashed black line; this is unavailable in practice, as nature provides but one realization. The true bias, again unknown to the data analyst, is denoted by the heavy black line. For small α (Fig. 7.2a), there is less variance in the 100 Monte-Carlo estimates of the bias. However, because the algorithm thereby provides heavier weight to past data, and because the true bias for the past data is seasonally dependent, there are systematic errors in those bias estimates; the maximum amplitude of the bias is typically

under-estimated and lags the true bias. The tradeoff made for this value of α has resulted in comparatively low variance amongst the bias estimates but high systematic error with respect to the true underlying bias. It is akin to a regression analysis with too few predictors (underfitting). For large α (Fig. 7.2d), much weight is provided to the most recent forecast deviation from the observations. The several recent observations are implicitly assigned heavier weight while the long-term mean is assigned less weight. This is akin to a regression analysis with too many predictors. The bias estimates change rapidly with each new daily update, and there is a much greater variety of bias estimates over the 100 independent simulations. The tradeoff made for this α has resulted in lower bias on average, but there is high sampling variability.

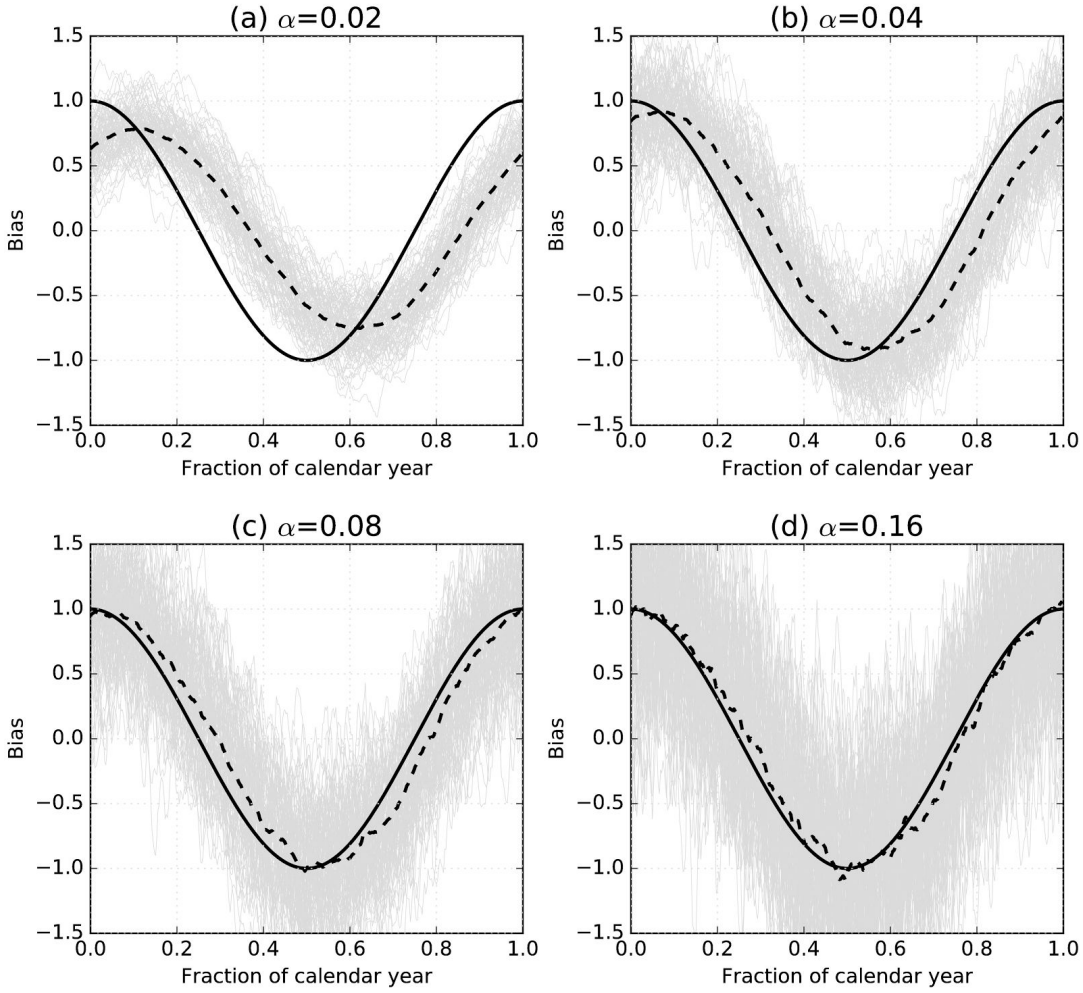

Figure 7.2. Illustration of the bias-variance tradeoff in statistical post-processing for the decaying-average bias-correction algorithm. Thin gray lines denote individual Monte-Carlo bias estimates using the decaying-average bias-correction algorithm. Dashed black line indicates the mean of the 100 Monte-Carlo bias estimates. Solid black line indicates the true underlying bias. Panels (a), (b), (c), and (d) show decaying average weights α = 0.02, 0.04, 0.08, and 0.16 respectively.

In practical weather prediction, again we have only a single set of observation data to work with, not 100 replicates, so the dashed lines in Fig 7.2 are never achieved. Were a data analyst to use this algorithm without modification, she would be faced with making a choice, adjusting α as to find an acceptable compromise between the bias and the variance based on seeing only a single one of the 100 thin gray lines in each of the panels of Figure 7.2. If she had developed some intuition that the biases were seasonally varying, she might choose to test the value of incorporating other predictors in a more sophisticated regression analysis, predictors such as $cos\,(2\pi J(t)/365)$ and $sin\,(2\pi J(t)/365)$. Why not do this? The decaying-average bias correction has one very appealing characteristic: very little data need be archived. Once the current forecast and observation have been used to update the bias, it can effectively be discarded for purposes of training. A longer time series of data would need to be stored to apply the more appropriate regression analysis and improve the bias estimates. While data storage was insignificant in this simple synthetic problem, if the method was applied to many variables on a high-resolution grid over a large area, the data storage demands might require the analyst's attention.

7.3. Training data issues for statistical postprocessing.

Consider now the characteristics of an ideal training data set, ideal not in the sense of providing perfect forecasts but rather ideal in that it serves nearly all the needs of the statistician.

- *The training data should span a long period of time, thereby providing multiple samples of the range of possible future environmental conditions*. This would provide enough samples to quantitatively estimate the probability of even relatively unusual events at each geographic location. Forecast errors are likely to be at least somewhat related to the local geographic peculiarities, including characteristics such as the terrain height, the terrain orientation, the vegetation, land-use, and soil type. With voluminous training data, models could be developed that incorporate any necessary additional predictors without overfitting.
- *Training data should be generated from the same ensemble prediction system in the training period as used for real-time predictions*. This makes the error characteristics of the forecast more consistent over time.
- *Real-time and retrospective forecast ensembles would have many members.* This permits estimates of the atmospheric uncertainty to be quantitatively estimated with modest sampling variability.
- *Error characteristics would not change radically over time*; the forecast errors from simulations 10 or 20 years past would be similar to those today.
- *Past analyses or observations used as predictand data would cover the same period as the forecasts*.
- *Past analyses or observations would be unbiased and of uniformly high quality*.
- *Observation or analysis data would be available for all the locations where post-processed guidance is desired*.

Unfortunately, less-than-ideal training data is the norm. Consider now the issues with predictor data (typically the ensemble forecasts) followed by the issues associated with predictand (observation, analysis) data.

a. Challenges in developing ideal predictor training data.

The ideal predictor data set can be computationally expensive to generate and archive, and may even be practically impossible to achieve perfectly. Let's presume that an operational implementation occurs every year, and that many past years of forecast data are desired. Computational costs will scale linearly with the number of ensemble members and the number of past "reforecast" cases; twenty years of reforecasts will be ten times more expensive than two years. Ideally, the model forecast data would be archived at the native model resolution, but if the forecast upgrade has double the horizontal resolution and twice as fine a temporal resolution, eight times more reforecast data must be stored when the model is changed, an increasing data-storage burden as the system is upgraded. If the statistical model development is occurring on another computing system, there are additional issues of data transfer to and storage on the computer system used for statistical development. While this may not be excessively burdensome if the statistical modeler is developing a regional post-processing system for one or two variables, it becomes an increasingly important issue to deal with if the system is intended to produce statistical adjustments for a wide number of variables over a large geographic region.

The ideal forecast data set would also generate the reforecasts' initial conditions using a consistent data assimilation system, the same one as used for the generation of the real-time forecasts' initial conditions. Most operational centres use a computationally expensive four-dimensional variational data assimilation technique (4D-Var; Courtier et al. 1994, Kalnay 2003), an ensemble Kalman filter (Hamill 2006, Evensen 2014), or hybridizations of the two (e.g., Buehner et al. 2013, Kleist and Ide 2015). Generating multi-year or even multi-decadal reanalyses to provide reforecast initial conditions may use computational resources that could otherwise be used for increasing the real-time prediction system's resolution or its ensemble size. The additional post-processed skill added by utilizing the extra training data must be evaluated relative to the additional skill generated from using a higher-resolution, more sophisticated real-time prediction system.

Perhaps to save the computational expense of regenerating reanalyses, the developers of a prediction system may choose to initialize reforecasts using a previously generated reanalysis based on an older version of the forecast model and assimilation system. This was the choice that was made with the recent NCEP Global Ensemble Forecast System (GEFS) reforecasts (Hamill et al. 2013). Prior to 2011, initial conditions were generated from the NCEP Climate Forecast System reanalysis (Saha et al. 2010). Subsequent to this, the forecast initial conditions were generated from the real-time data assimilation system, which underwent various changes that affected initial condition characteristics. Figure 7.3, from Hamill (2017), shows

that the character of short-range temperature and dew point analyses changed over that period with respect to an unchanging reanalysis developed at the European Centre for Medium-Range Weather Forecasts (ECMWF; Dee et al. 2011). Forecasts inherit this initial-condition bias to some extent, so the statistical character of the forecasts were not homogeneous before vs. after 2011. The practical impact of this is degraded statistically post-processed products after 2011 if they were trained with forecast data prior to 2011.

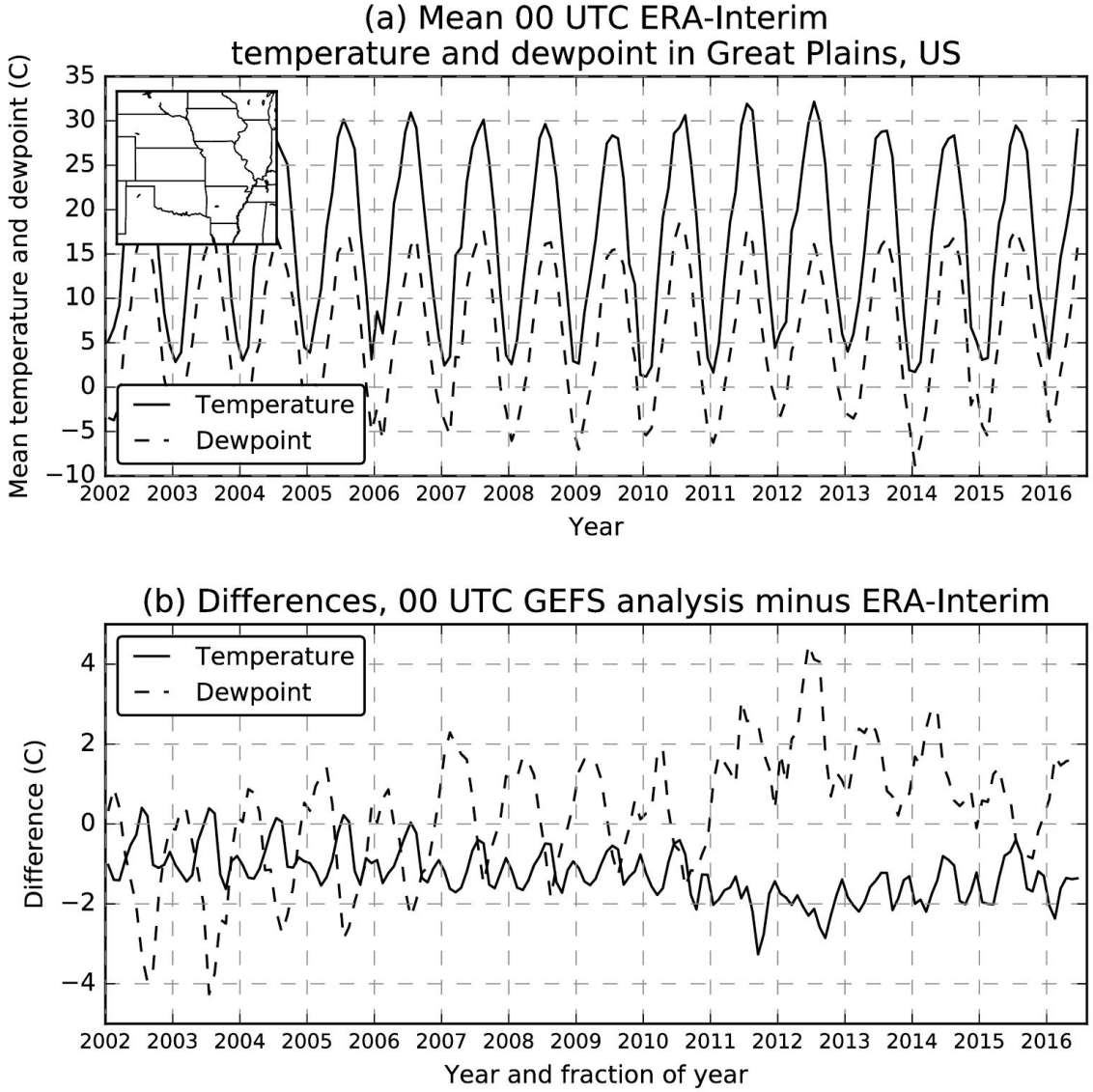

Figure 7.3: (a) Time series of mean temperatures at 00 UTC from ERA-Interim reanalyses for area covered in map inset. (b) Time series of mean differences at 00 UTC between the temperature of the GEFS initial analysis and the ERA-Interim analysis for temperature (solid curve) and dewpoint (dashed curve).

Even if the computational and storage resources are set aside for the generation of multi-decadal reanalyses and reforecasts that are consistent with the operational prediction

system, the observing system that provides input to the reanalysis system may have changed dramatically over the reanalysis period. In the last few decades, assimilation systems have begun to assimilate more and more satellite data, including microwave radiances (McNally et al. 2006), infrared radiometer data (Collard and McNally 2009), cloud-drift winds estimated from high-resolution satellite imagery time series (Velden et al. 2005), aircraft temperatures (Benjamin et al. 2010), scatterometer estimates of ocean surface winds (Bi et al. 2011) and radio occultations (Anthes et al. 2008). These have increased the accuracy of analyses and reanalyses in recent years. Because of these changes, even with current state-of-the-art assimilation methods it is not possible to generate a retrospective forecast for a date in the distant past with expected errors as small as they are for current forecasts (Dee et. al. 2011, Fig. 1).

b. Challenges in gathering/developing ideal predictand training data.

Training against gridded analyses is often desired, for many users need gridded post-processed guidance, and this is a straightforward way to achieve this. Unfortunately, some of the characteristics of the ideal analyses outlined above are difficult to achieve. First, a long time series of analyses can be computationally expensive if generated with modern data assimilation methods such as 4D-Var, the EnKF, or hybridizations. It also requires synthesis of all available observations and massive storage of the resulting data. This may make reanalysis generation impractical for some prediction centers. Were the statistician to use the operational analyses produced in real time, these analyses would likely vary in quality and bias, reflecting both the changing nature of the observing system and the changes in the data assimilation and forecast system.

Why should one expect the analysis bias to vary over time? Presumably the analyses or reanalyses are generated by adjustment of first-guess (background) forecasts to newly available observations. Then the observations and the background should be unbiased in order for the assimilation procedure to produce the unbiased analyses desired for postprocessing. While technology for adjusting the observations to reduce bias (e.g., Auligné et al. 2007) is now common, and while approaches to adjust the background to be unbiased have been proposed (e.g., Dee 2005) if not widely used, complete removal of bias from data assimilation information sources is still problematic. Hence many analyses should be expected to have bias.

Illustrations of analysis bias are shown in Figs. 7.4 and 7.5. Figure 7.4 shows the time-averaged spread (standard deviation about the multi-analysis mean) of 2-meter surface temperatures between four different prediction centers. Spreads are calculated for each day and then averaged over the year. Analyses were interpolated to a 1-degree grid before display and were taken from the TIGGE archive (Bougeault et al. 2009, Swinbank et al. 2016). Time-averaged analysis spreads exceeding 1° C are common, with many regions, especially in mountainous and polar regions, with much greater spread. If we examine the time series of analyses at a particular location (Fig. 7.5), here in the central Amazon river basin, we see that the differences are not random; some analysis systems are systematically colder, others

systematically warmer than the average. Choosing one (e.g., NCEP) and training against this analysis is thus likely to result in post-processed guidance at this location that has a warm bias (presuming the multi-center mean is more realistic). Note that differences between analyses for upper-air variables may not be as pronounced (Park et al. 2008), as near-surface variables are especially challenging to predict given that many of the relevant processes (boundary layer, surface layer, land surface) are treated through parameterizations, i.e., approximations of the sub-grid scale effects upon the resolved scales (Stensrud 2007).

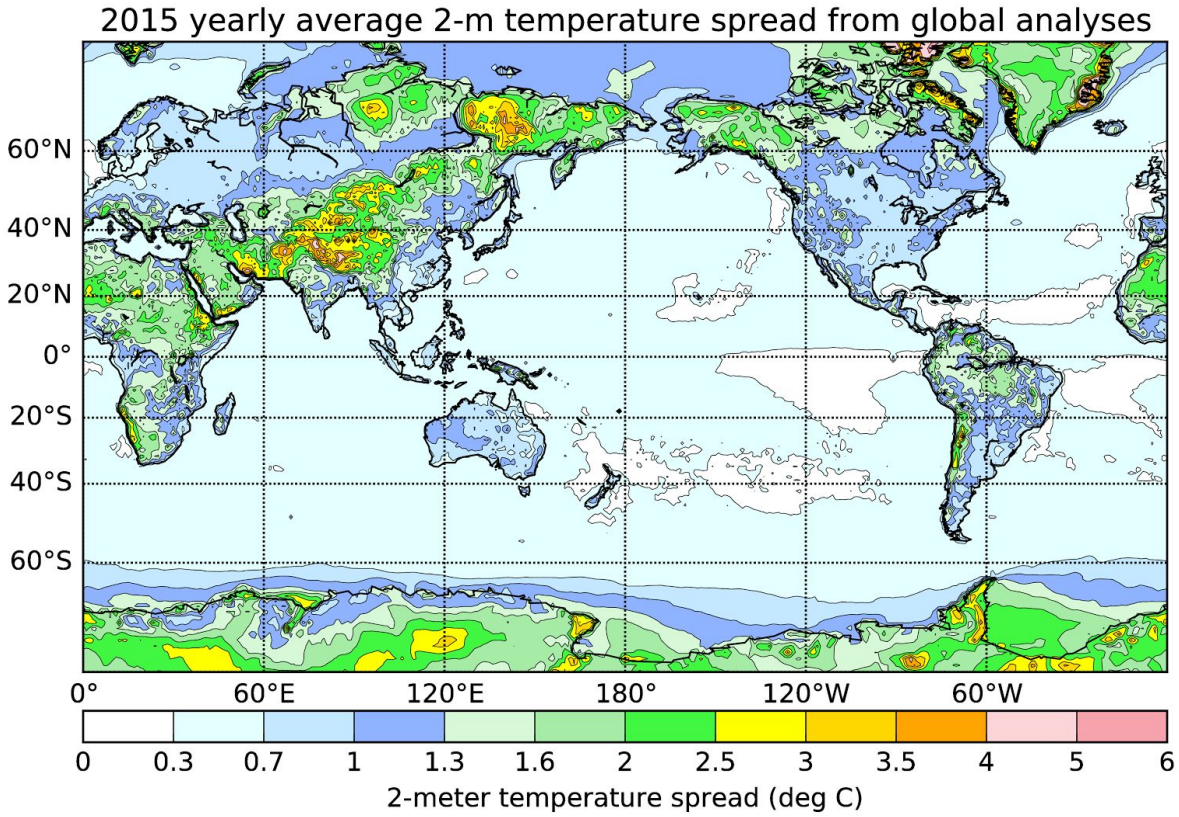

Figure 7.4: 2015's yearly average of the daily spread of 00 UTC 2-meter temperature analyses. Data for each analysis system was extracted on a common 1-degree grid via European Centre for Medium-Range Weather Forecasts (ECMWF) TIGGE data portal (Bougeault et al. 2009). Analysis systems used here were National Centers for Environmental Prediction (NCEP), Canadian Meteorological Centre (CMC), the UK Met Office, and ECMWF.

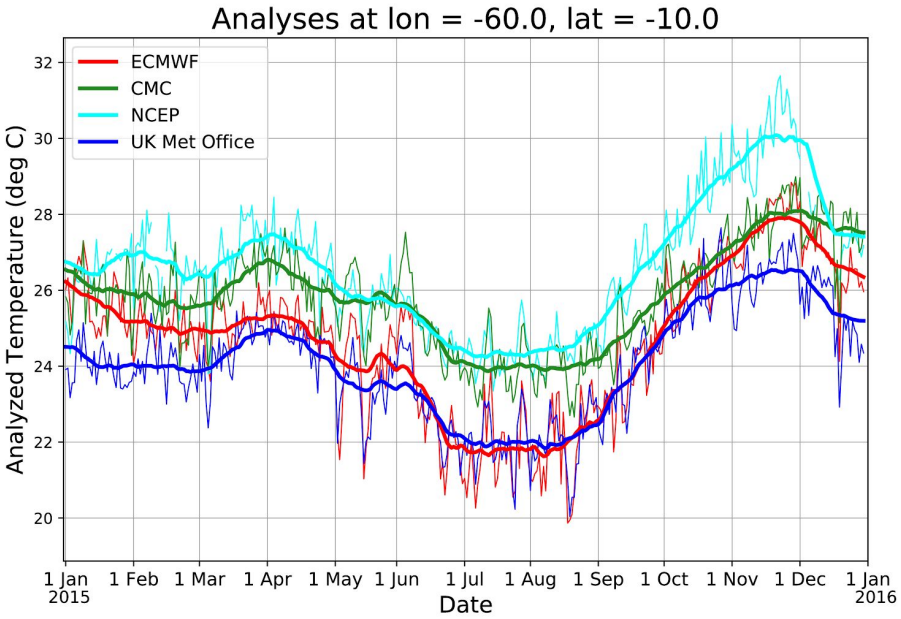

Figure 7.5: Raw (thin lines) and +/- 15 day smoothed (thick lines) time series of 2-meter surface temperature analyses at 00 UTC from four different global data assimilation systems for a location in the Amazon river basin.

Given the challenges with training against analysis data, even if gridded products are preferred, might it be preferable to directly use station data? Training against observations provides more site-specific, downscaled information specifically at the observation site (Vannitsem and Hagedorn 2011). However, if information is also desired at other nearby locations, spatial modeling is necessary. Several such techniques have been developed. These include the procedure of Glahn et al. (2009), where postprocessing is first performed at stations and then interpolated to a grid. Scheuerer and Büermann (2014) proposed a strategy, further developed by Dabernig et al. (2017) and Stauffer et al. (2017) where climatological characteristics are interpolated to the grid and removed from both forecasts and observations, so that all locations within a region can be postprocessed simultaneously.

While these methods avoid training against analyses that contain bias, there are disadvantages to such approaches. For example, in-situ observation locations are commonly sparse over bodies of water and in mountainous and less-populated regions. Given that commonly desired variables like temperature, wind speed, and precipitation amount may vary with elevation and vary from land to ocean, such statistical interpolations from observation locations to the output grid may produce lower-quality grids in such regions than are desirable. Analyses produced through data assimilation procedures, despite their contamination by bias, will often have useful information in areas devoid of in-situ observations. This is both because they use other sources of data such as satellites and radars, and because the model first-guess field (background) is effectively a repository of information accumulated from the assimilation of earlier observations.

7.4. Proposed remedies for practical issues in statistical post-processing.

a. Improving the approaches for generating reforecasts.

Let's assume for the moment that the director of a prediction center has made a determination that postprocessing is an important step in the production of forecasts, and that training sample size is of sufficient importance that some computational resources must be set aside for reforecasts. Let's also assume for the time being that reanalyses that are similar in quality to the real-time analyses have already been generated. The director of the prediction center has perhaps indicated that the number of reforecasts that can be generated without unduly affecting the implementation of other model improvements is not extravagant, perhaps limited to running the ensemble system retrospectively spanning four years of training data with, say, 5 members. With these limits established, other configurations are possible. Two years could be spanned with 10-member reforecasts at the same computational expense. This would decrease the range of weather scenarios covered, but ensemble spread estimates for each case would be improved. Also, twenty years of reforecast data could be generated by generating a 5-member reforecast every fifth day (Hamill et al. 2004). A regular, every-nth-day sub-sampling procedure may be nearly optimal for some variables but not for others. Suppose the most important intended application of the reforecast data set was the statistical postprocessing of heavy precipitation. In such a situation, weather-dependent procedures for determining the dates to generate reforecasts might improve the postprocessing of heavy precipitation. For example, the probability that one should generate a reforecast for a particular past date could depend on, say, the likelihood that a prior-generation reforecast was predicting much heavier than average precipitation amounts in areas of particular interest[2]. When determining a list of case days to reforecast, a day with 20 percent probability of heavy precipitation would be twice as likely to be selected as a day with 10 percent probability. Were reforecasts selected based on probability of heavy precipitation occurring somewhere in a large region (say, the contiguous US), one would expect that at any specific point, there would still be many samples of more ordinary weather, and the accuracy postprocessing algorithm would not be degraded for more common events.

Are there general principles that might guide reforecast configuration? Such decisions should be informed by the intended application(s). If the primary application is for sub-seasonal forecasting where the forecast is more affected by boundary conditions such as sea-surface temperature and soil moisture than by the initial atmospheric state, then a reforecast data set spanning a wider range of climate states is desirable; twenty years every fifth day is preferable to four years every day. If shorter-term probabilistic precipitation postprocessing is of greater

[2] What should *not* be done is to select cases on the basis of observed or analyzed heavy precipitation. In such a situation, the training data would be biased toward the occurrence of heavy precipitation events, and the post-processing technique applied to the real-time forecast would likely over-predict the precipitation.

interest, a weather-dependent sampling strategy that generates reforecasts on days where heavy precipitation is more likely is desirable.

Are there principles to determine the tradeoff of length of reforecast vs. ensemble size? Again, this may depend on the intended application. For products like the extreme forecast index (LaLaurette 2003, Petroliagis and Pinson 2014) that use the reforecasts to determine how unusual today's forecast is relative to the ensemble reforecast climatology, ECMWF's experience (Vitart et. al. 2014) has shown that the product performance is improved with more members. On the other hand, for many statistical post-processing applications, it can be more helpful to have a greater number of individual weather events than a larger ensemble. The primary improvement skill in postprocessing is a more commonly a result of correcting errors in the mean state than adjusting the spread, and a wider range of weather scenarios permits more appropriate state-dependent corrections.

How might one address the changing statistical quality of the reforecasts over time due to observing network changes? Past experience has shown (Uppala et al. 2005, Fig. 14, vs. Hamill et al. 2013, Fig. 1; also Dee et al. 2014, Fig. 1) that the more advanced the data assimilation and prediction system, the higher the overall quality and more uniform the statistical characteristics of past vs. current forecasts. Hence, regular production of reanalyses with the most up-to-date system is the most straightforward way to address this, however impractical it may be computationally. Assuming one does not have regularly generated reanalyses of uniformly high quality, other options might include weighting the training samples to be inversely proportional to their expected error variance, as is done for example in weighted least-squares regression. If reanalyses are not available but some computational resources have been set aside for reforecasts, perhaps judicious use of a different reanalysis for initialization may prove useful. In the recent past, prediction centers without their own reanalyses have explored creating reforecasts through a modified initialization with another center's reanalyses, adjusted near the surface to reflect the climatology of the own center's land-surface scheme (Boisserie et al. 2016, Lin et al. 2016).

Should a prediction center not have reanalyses readily available but believe them to be necessary, their generation may be the most expensive and time-consuming part of the process. Suppose we want to generate a 5-member reforecast every third day to +30 days lead. Every third day we have thus generated 150 member days of reforecasts. Say now that an 80-member ensemble data assimilation approach is used, stepping the 80 members forward six hours, producing an updated analysis, and repeating the process. The computational expense of merely generating the background ensemble of forecasts for the data assimilation over the same 3-day period is 80 x 3 = 240 member days. The computational expense of the analysis update step is roughly the same order of magnitude. Because of the large computational expense and labor involved in reanalysis generation, reanalyses are thus typically generated once or twice a decade at some prediction centers (e.g., ECMWF, the US NWS, and Japan Meteorological Agency) or not at all for many others.

In the end, is the computation of reforecasts, and perhaps reanalyses, a necessary precursor for effective statistical postprocessing? It likely depends on the intended application. Previous experience (e.g., Hamill et al. 2006, Scheuerer and Hamill 2015) have shown that for rare events like heavy precipitation, the enlarged sample size afforded by many reforecasts improves the post-processing skill significantly. Another application that is greatly improved by reforecasts is the post-processing of subseasonal forecasts. At these leads the noise due to chaos and model error is large and the detectable signal is small; large samples are helpful to extract the small amount of signal amongst the bath of noise (Ou et al. 2016). Reforecasts also provide the large sample sizes that can be important for validation of rare, extreme events, such as heavy precipitation events leading to floods. For other applications such as short-term temperature calibration (Hagedorn et al. 2012) or basic probability of nonzero precipitation amount forecasting (Hamill et al. 2017), it may be possible to work around some of the issues related to the short training data set, as discussed in the next subsection.

b. Circumventing common challenges with shorter training data sets.

Suppose lengthy reforecasts are a practical impossibility and one must make do with a much shorter time series of forecasts for training. What procedures may be practical in such a circumstance? For some variables such as surface temperature, past experience shows that some benefit can be obtained with simple approaches such as the decaying-average bias correction discussed in section 7.2. This is because bias commonly has a large systematic component, especially at short leads, and hence yesterday's forecast bias provides useful predictive information on today's bias. For other variables of interest such as precipitation, the past few days or weeks or even months may not provide a wide enough variety of precipitation events to achieve major improvements, especially if the post-processing method is applied independently from one location to the next. The limited sample size requires consideration of other approaches.

An obvious candidate is the supplementation of training data using information from surrounding regions (e.g., Allen and Erickson 2001, Mass et al. 2008). Figure 7.6 provides evidence for why one should be judicious with such approaches. Here, GEFS reforecast data (Hamill et al. 2013) and climatology-calibrated precipitation analyses (CCPA; Hou et al. 2014) were used for the Dec-Jan-Feb 2002-2015 period to populate cumulative distribution functions (CDFs) of 24-h accumulated precipitation at two nearby locations along the Oregon-California border in the northwest US. Suppose one were one to apply a quantile-mapping procedure (Hopson and Webster 2010, Voisin et al. 2010, Maraun 2013) to address the conditional bias of the forecasts. For example, perhaps the coastal location's training data is supplemented with the training data from the inland location. The conditional forecast biases of moderate precipitation at these two locations are opposite in sign. Along the coast, precipitation is under-forecast, while slightly inland it is over-forecast. Applying quantile mapping to the coastal location using the inland supplemental training data would likely produce a worse adjusted forecast than were the data kept separate.

Perhaps the concept of using supplemental training data is conceptually sound provided one is careful with what supplemental data is used. A more advanced selection procedure for supplemental data was recently demonstrated in Hamill et al. (2015, 2017) and a similar approach was discussed in Lerch and Baran (2017). For each grid point where a post-processed precipitation forecast was desired, a number of supplemental locations were determined based on similarities of climatology and geographical characteristics such as terrain height and hillslope orientation, with the presumption that many precipitation biases are related to the simplified representation of the terrain characteristics in the numerical model. Training data at the original location was supplemented with the data at these additional locations, with subsequent improvement to the post-processed guidance.

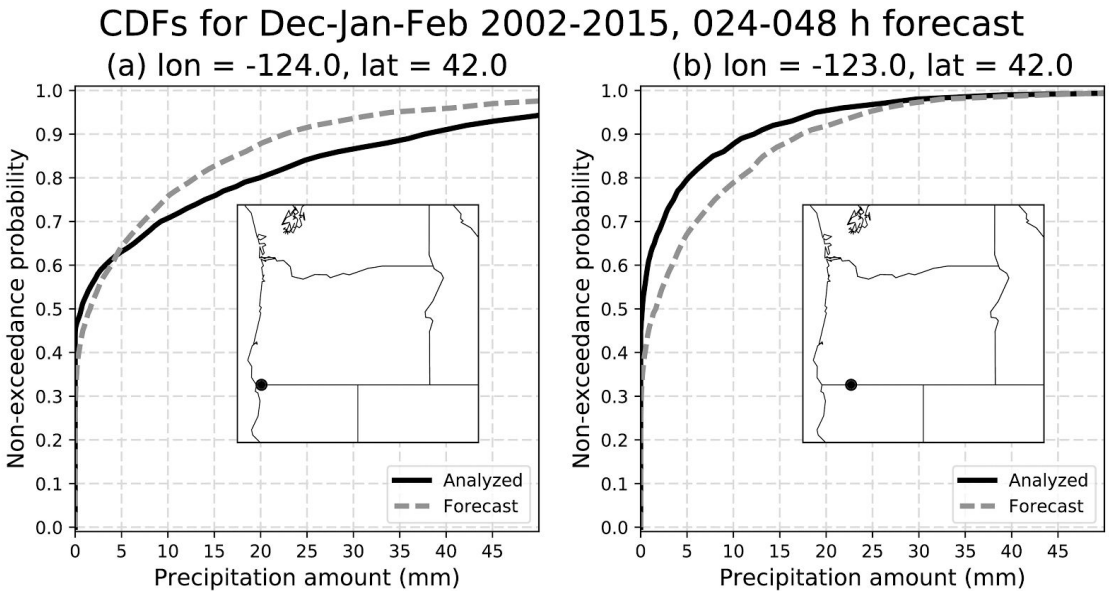

Figure 7.6. Illustration of regionally dependent differences between forecast and analyzed precipitation, here at two locations in the western US. Analyzed CDFs from CCPA analyses are shown in the heavy black lines, while GEFS member forecast CDFs are shown in with heavy dashed grey lines. The two locations are denoted by the two black dots in the inset maps.

Another issue that should be considered when working with short training data sets is the possibility of seasonally dependent bias, as with the example of Fig. 7.2. Suppose for practical considerations one must use only the last month or two of forecast and observed/analyzed data for training. Again, focusing on the statistical postprocessing of precipitation, consider forecast and analyzed CDFs again from a reforecast data set for three sequential months (Fig. 7.7). The differences between forecast and observed at moderate amounts for this location in southwest Iowa (US) change from a relatively neutral bias at higher precipitation amounts in February to a slight under-forecast in April and a more noticeable under-forecast in June. If the model hasn't changed in more than a year, the most straightforward approach to deal with this is to use additional training data from the same season but from the previous year.

Suppose a post-processing application requiring many years of retrospective forecasts. Consistent reforecasts are unavailable, but an archive of past forecast guidance from the operational model is available. Optimistically, perhaps only the mean bias changes with model version. In this case an indicator (or "dummy") variable (Neter et al. 1990, Chapter 10) may suffice to permit use of multiple model versions in the training data were regression-type approaches used for postprocessing. Perhaps the regression relationship changes in other ways, leading to the need for a larger set of predictors and interactions for each model version. In such a case one must be aware of the potential for overfitting.

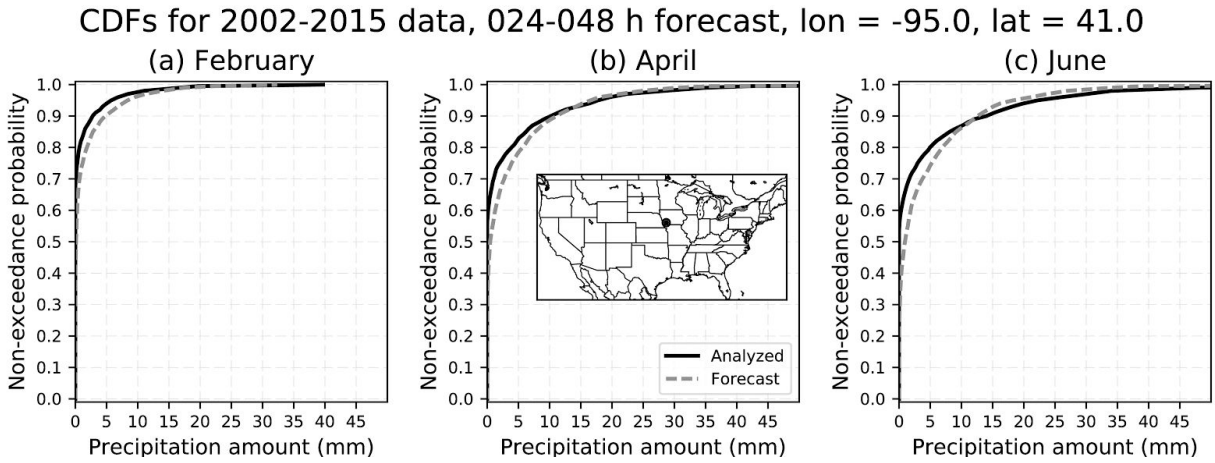

Figure 7.7. Cumulative distribution functions of 24-h accumulated CCPA-analyzed (heavy black curve) and GEFS member's reforecast (dashed grey curve) precipitation for three months for a location in SW Iowa of the US. Curves were generated using 2002-2015 data. (a) February, (b) April, and (c) June.

c. Substandard analysis data.

If training against analyses is an imperative for postprocessing, and if the analyses have systematic errors as previously discussed, one possible but time-consuming approach for remediating these errors is to improve the data assimilation system that generates the analyses. Meteorological statisticians in the US National Weather Service (NWS) have requested such an improvement from the NWS data-assimilation system developers. The NWS wishes to perform postprocessing against high-resolution gridded analyses. The current NWS high-resolution analysis system (e.g., de Pondeca et al. 2011) produce analyses with bias and higher-than-ideal errors, especially in the mountainous western US. Hence, the NWS is providing resources for the improvement of this analysis system. Unfortunately, the analysis variables that are of greatest interest (temperature, wind, precipitation, and so forth) are often the most difficult to improve. The accuracy of estimating these variables depends on faithfully depicting the atmosphere interacts with the land surface, with all its heterogeneities, its physical complexity, and its poorly observed soil state. Further, significant errors in the depiction of clouds ubiquitous in prediction systems may contaminate the model estimates of downward solar

radiation that largely determine the resulting surface sensible and latent heat fluxes back to the atmosphere.

Another reason that high-quality analyses suitable for postprocessing are challenging to generate is that the postprocessing requirements for analysis data may be somewhat different than requirements for forecast initialization. For postprocessing, accuracy, lack of bias, and relevant spatial detail are of paramount importance. For forecast initialization, an analysis that leads to an accurate and stable forecast is paramount. Introducing, say, the use of the actual terrain heights in the system rather than a smoothed version may produce somewhat more realistic analyses but radically poorer predictions.

While the direct improvement of gridded analyses is desirable, major improvements will take time to achieve, and some useful analysis data may be needed right away. Here are some suggested guidelines for use of analysis data: (a) if multiple analyses are available (e.g., Fig. 7.5), then consider training and verification against some linear combination of the available analyses. The underlying hypothesis is that the different systems may have somewhat independent biases, and a mean will have a more accurate estimate than any one individually. (b) Consider approaches that leverage station data but implicitly produce a gridded product, as in Glahn et al. (2009), Kleiber et al. (2011ab), Scheuerer and König (2014), Scheuerer and Möller (2015), and Stauffer et al. (2017).

7.5 Case study: postprocessing to generate high resolution probability of precipitation from global multi-model ensembles.

A practical example of a challenging problem in statistical postprocessing is now presented, illustrating some of the tradeoffs discussed above and the choices made in the development a pre-operational post-processed product in the US.

Several years ago, the US NWS instituted a statistical postprocessing initiative, the "National Blend of Models" or more simply the "National Blend." Many of the worded weather forecasts produced by the NWS are automatically generated from gridded fields of temperature, winds, precipitation, and so forth. Forecasters at several dozen NWS offices typically provide manual modifications to centrally produced model guidance grids. When one views the synthesized product nationally, abrupt discontinuities are sometimes evident at the boundaries between two weather forecast offices' area of responsibility. The National Blend intends to provide statistically post-processed guidance from multi-model ensembles to the forecasters, guidance of such quality and reliability that the need for manual editing is much less necessary. The intent is to improve forecast consistency as well as quality.

The following case study illustrates a statistical postprocessing method in development for 12-h probability of accumulated precipitation (POP12) in the National Blend. In the US, nonzero precipitation is defined as the event of ≥ 0.254 mm, here during the 12 h period. The ultimate desired guidance is a 2.5-km grid of POP12 over the contiguous US (CONUS) and

adjacent coastal waters, with likely future extension to a full probabilistic quantitative precipitation forecast. For the initial technique development described here, POP12 is produced and validated on a ⅛-degree grid, which is equivalent to approximately 10.6-km grid spacing at 40° N latitude. The data inputs consist of global deterministic and ensemble forecast guidance and ⅛-degree Climatology-Calibrated Precipitation Analyses (CCPA; Hou et al. 2014). CCPA has been available since 2002, making it useful for determining a precipitation climatology as well as postprocessing model training and validation. Unfortunately, CCPA data does not cover adjacent coastal waters, one of the tradeoffs made in this application.

In this initial development stage, only two ensemble systems are used for medium-range POP12 forecasts, the US National Centers for Environmental Prediction (NCEP) Global Ensemble Forecast System (Zhou et al. 2016), and the Canadian Meteorological Center (CMC) Global Ensemble Prediction System (Gagnon et al. 2014, 2015). Hereafter, these are referred to as the "NCEP" and "CMC" ensembles. Each ensemble system provides twenty ensemble-member forecasts at a resolution coarser than ⅛ degree. Single deterministic control forecasts are also used from each center. In the future, US Navy global ensemble will be used as well, though these data are not part of this study. For more on the multi-center US and Canadian ensemble, see also Candille (2009).

The postprocessing method demonstrated here is the approach described in Hamill et al. (2017), which provides greater detail on the methodology and the rationale for its use. The methodology combines a variety of established algorithms, chosen for their suitability to short training data sets and multi-model ensembles. The approach is also readily extensible to full probabilistic quantitative precipitation forecasting in the future. The algorithmic approach includes five general steps: (1) Populate forecast and analysis CDFs of precipitation using the last 60 days of data. To increase the training sample size, CDFs at a particular grid point for a particular ensemble member were populated not only with training data for that grid point, but also from data at that grid point's predefined supplemental locations (ibid). (2) Quantile map each ensemble member using the forecast and analyzed CDFs. This ameliorates conditional bias and applies an implicit statistical downscaling. (3) Dress each quantile mapped ensemble member with random noise to correct for remaining problems with underdispersion. (4) Generate probabilities from weighted, dressed ensemble members. (5) Smooth the resulting POP field.

The first step in the real-time data processing is populating the forecast and analyzed CDFs with the prior 60 days' data. As noted earlier, postprocessing of precipitation can be very difficult with small training sample sizes, but this is ameliorated here (e.g., Fig. 2 from Hamill et al. 2008) by supplementation of the training data with data from other locations with similar terrain and precipitation characteristics. Specifically, for each output ⅛-degree grid point, a set of other grid points, or "supplemental locations" are identified, and then the forecast and analyzed CDF for this grid point is populated with data from the original grid point and with data from the supplemental locations. Such an approach is often preferable to enlarging sample size by combining training data from radically different seasons that often have different

conditional biases (Fig. 7.7). Also, should the prediction system change versions and have different biases for different system versions, older model data will be aged out by the end of 60 days as well, limiting the potential duration of degraded product quality.

Figure 7.8 shows predefined POP12 supplemental locations for several selected grid points in the US during the month of April. Though supplemental locations are shown only for six grid points in this figure, supplemental locations are defined for every ⅛-degree grid point in the CONUS and the Columbia river basin of Canada. The supplemental locations were chosen based on similarity of the CCPA precipitation climatology during the 2002-2015 period, terrain height and aspect, and physical separation between grid points. The rationale was that the model's location-dependent systematic errors of precipitation were related to the precipitation climatology in part, and also to the smoothed representation of the terrain relief in the coarser-resolution numerical models (Fig. 7.6). Supplemental locations were prevented from being too close to each other so that samples had more independent error characteristics. For more information on the supplemental location algorithm, see Hamill et al. (2017).

The next step in the real-time processing is to apply quantile mapping to each ensemble member using the CDFs generated in the previous step. Quantile mapping is illustrated in Fig. 7.9. Presuming a CDF has been generated for a grid point and ensemble member, we determine the forecast amount (here, 4 mm) and its non-exceedance probability (here ~0.895). The analyzed amount associated with the same non-exceedance probability is determined (3 mm), and the forecast amount is adjusted to this value. The procedure is repeated for each output grid point and each ensemble member. This procedure permits the conditional bias to be mitigated. Should the analysis be on a more finely spaced grid with more detail, then the algorithm is also implicitly performing a statistical downscaling.

Other procedures such as Bayesian Model Averaging (Raftery et al. 2005, Sloughter et al. 2007) adjust for forecast bias through regression approaches. As noted in Wilks (2006) and Hodyss et al. (2016), when regression equations are applied to ensemble members under situations where there is little predictive relationship between forecast and analyzed, the ensemble members are regressed toward the mean analyzed value, resulting in an ensemble with a reduction in spread. The reason ensembles are generated in the first place is to provide situational estimates of the forecast uncertainty, and raw ensembles are commonly under-spread; regression of each member can make a bad problem worse. It is for this reason that quantile mapping was preferred over regression; the ensemble spread is less affected, and the subsequent dressing step discussed below has less "work" to do.

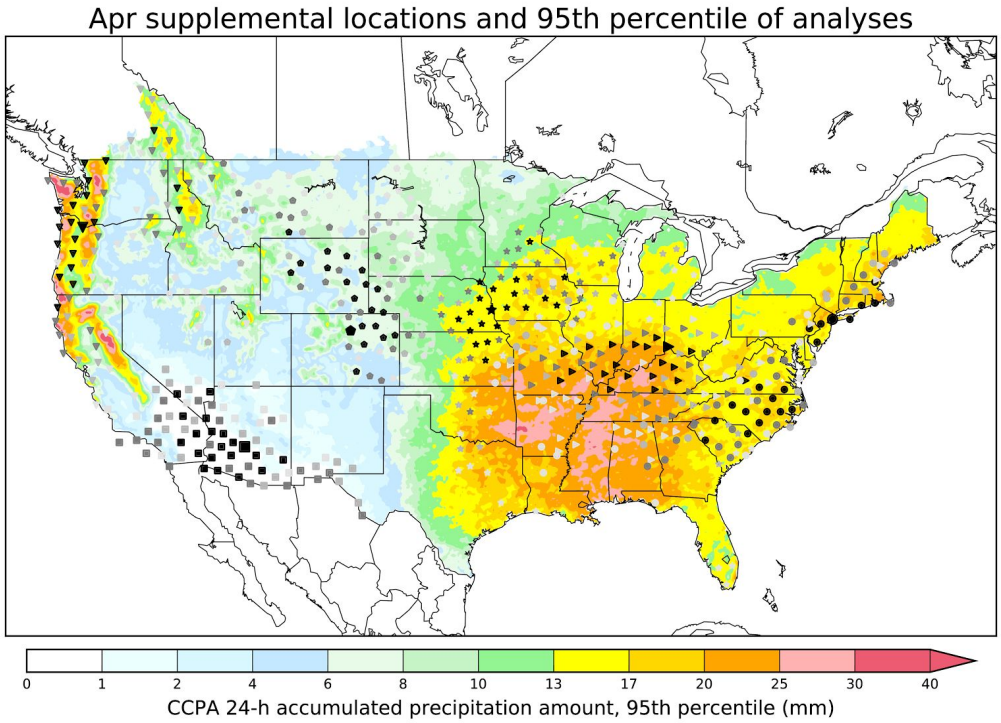

Figure 7.8. Illustration of supplemental locations for the month of April. Larger symbols denote the locations for which supplemental locations were calculated (roughly Portland, OR; Phoenix, AZ; Boulder, CO; Omaha, NE, Cincinnati, OH, and New York City, NY). Smaller symbols indicate the supplemental locations. Darker symbols indicate a better match, lighter symbols a poorer match. The colors on the map denote the 95th percentile of the 24-h accumulated precipitation amounts for the month, determined from a climatology of 2002-2015 CCPA data. Reprinted with permission from Hamill et al. (2017).

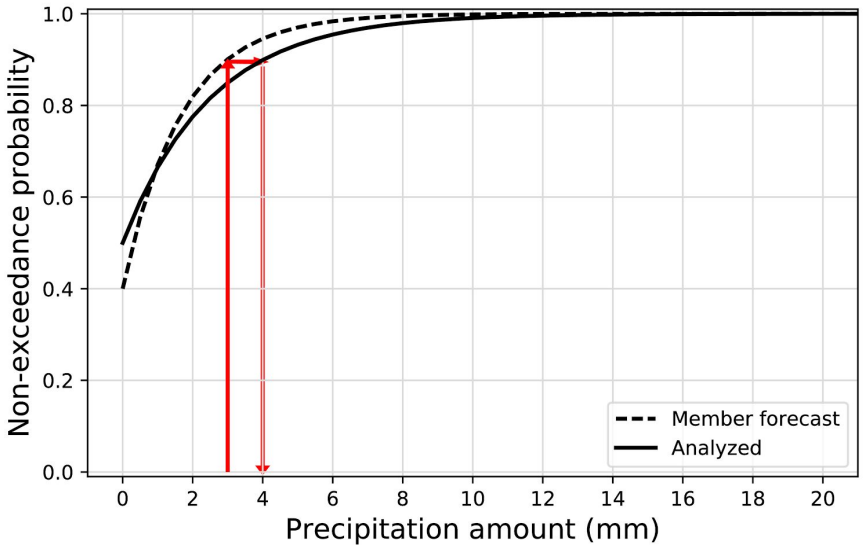

Figure 7.9. Illustration of the deterministic quantile mapping procedure applied to ensemble members. Forecast and analyzed distributions adjust the raw forecast to the analyzed value associated with the same cumulative probability. Grey arrows denote the mapping process.

An additional feature of the POP12 quantile-mapping procedure is briefly described. Should probabilities now be determined from the 42 ensemble members, one would expect some unreliability and loss of skill in part from the relatively modest size of the ensemble (Richardson 2001). To ameliorate this and to deal with overconfidence in ensemble systems in the positioning of precipitation features, the quantile mapping utilizes not just the forecast at the grid point of interest, but also quantile maps forecasts from surrounding grid points. In particular, that grid point and eight nearby points are used as input to the quantile mapping. For each nearby point, the forecast CDF used is the one associated with that nearby grid point, while the analyzed CDF is the one associated with center grid point of interest. In this way, forecasts from surrounding locations, even if they are in mountainous terrain with different climatologies, are mapped to be consistent with the analyzed distribution at the interior point. This process then provides a nine-fold larger ensemble, minimizing errors attributable to finite ensemble size. See Hamill et al. (2017) for more rationale and figures illustrating the process, and see Scheuerer and Hamill (2015) for another application of a similar procedure.

At this stage of the procedure, a ninefold larger quantile-mapped ensemble has been produced for each output grid point with location-dependent conditional bias reduced. There may be remaining errors such as insufficient ensemble spread, but they are assumed to be independent of location, though likely dependent on amount. The errors of the quantile-mapped members are also assumed at this point to be exchangeable; forecast member 1's quantile-mapped error statistics are assumed the same as forecast member 42's. Inspired by Fortin et al. (2006), a best-member dressing procedure is now applied. Each quantile mapped member's value is perturbed with random, normally distributed noise with mean zero and standard deviation $0.2 + 0.3 \times$ the quantile mapped value. Dressed values below zero precipitation are reset to zero. While this procedure is ad-hoc, it was informed by other experiments (not shown) where Gamma dressing distributions were objectively fitted.

The next step of the procedure is relatively straightforward. POP12 is estimated from the ensemble relative frequency, i.e., if 30 percent of the members have precipitation above the POP12 threshold of 0.254 mm, the probability is set to 30%.

The final step improves the visual appearance of the forecasts. There are small-scale variations in the POP12 field that are attributable to finite sample size and the application of random dressing noise. However, not all small-scale variations are noise. Over mountainous regions, small-scale variations may reflect orographically enhanced precipitation. Accordingly, we do a final Savitzky-Golay (Press et al. 1992) smoothing of the POP12, with more aggressive smoothing in flatter areas and less smoothing in areas with more variations in elevation. There are also procedures to taper the probabilities from their calibrated values to raw multi-model ensemble values beyond the borders of the US. See Hamill et al. (2017) for specifics.

Figures 7.13 and 7.14 provide a case study of how +60 to +72 h POPs are changed through each stage of the postprocessing. Figure 7.10(a) shows the verifying precipitation analysis, with heavy precipitation in the central US, from Texas north to Kansas. There was

also a smaller, north-south band northward from Mississippi to Wisconsin and Michigan. Scattered lighter precipitation occurred the Rocky Mountains of Colorado and Wyoming. The raw NCEP ensemble in Fig. 7.10(b) was overconfident of precipitation in many regions where no precipitation occurred, including in Arkansas, N. and S. Carolina, and in the northwest US. The raw CMC ensemble in Fig. 7.10(c) also forecast elevated POP12 in the northwest US and probabilities above 80 percent in a wide swath of the central US. As expected, the raw combined data shown in Fig 7.10(d) portrays intermediate probabilities between these two. Figure 7.11(a) presents the results if only quantile mapping were applied using the center of a 3 x 3 array of grid points, i.e., not using the surrounding data nor thereby increasing the ensemble size ninefold. The quantile mapping reduces the areal extent with low but nonzero POP12s in the western US, adjusting for the model tendency to over-forecast light precipitation amounts. The areal extent with very high POP12 in the central US was also decreased, with many areas with 95% or greater probability reduced to ~80 percent. The effects of statistical downscaling are also evident in the western US, where, for example, POP12 was decreased in Oregon, but much less so along peaks of the Cascade range, so they now appear as local maxima. The nonzero POP12 in N. and S. Carolina was reduced to near zero in many locations. When quantile mapping included the 3 x 3 array of surrounding points (Fig. 7.11(b)), there were many grid points whose probabilities were lowered further, and the probabilities east of the Rocky Mountains had a more smooth characteristic. The dressing algorithm (Fig. 7.11(c)) also made the forecasts less sharp in general, but they add some undesirable small-scale noise. This is largely diminished in the final product, shown in Fig. 7.11(d). This final product still has deficiencies; for example, the final POP12 has a single north-south band of higher probabilities in the central-southern US, while the observed precipitation had two bands. Ideally, the postprocessing would have reduced probabilities to zero throughout most of the western US, as that region was analyzed as dry. Nonetheless, the overall product exploits the diversity in the positioning of precipitation between the two systems, and it reduced POP12 in many regions with high raw probability but no occurrence of precipitation.

The various steps of the algorithm each contribute to the improvements in reliability and skill. Figure 7.12 shows reliability diagrams at the +60 to +72 h lead time for dates from 1 April to 6 July 2016. Quantile mapping using the center point only improves reliability and skill substantially, but the use of the 3 x 3 stencil improves it a bit more. Application of the dressing algorithm improves the reliability and skill a bit more. Smoothing does little to the skill, despite the improvement in the visual appearance of the forecasts (Fig. 7.11 (c)-(d)).

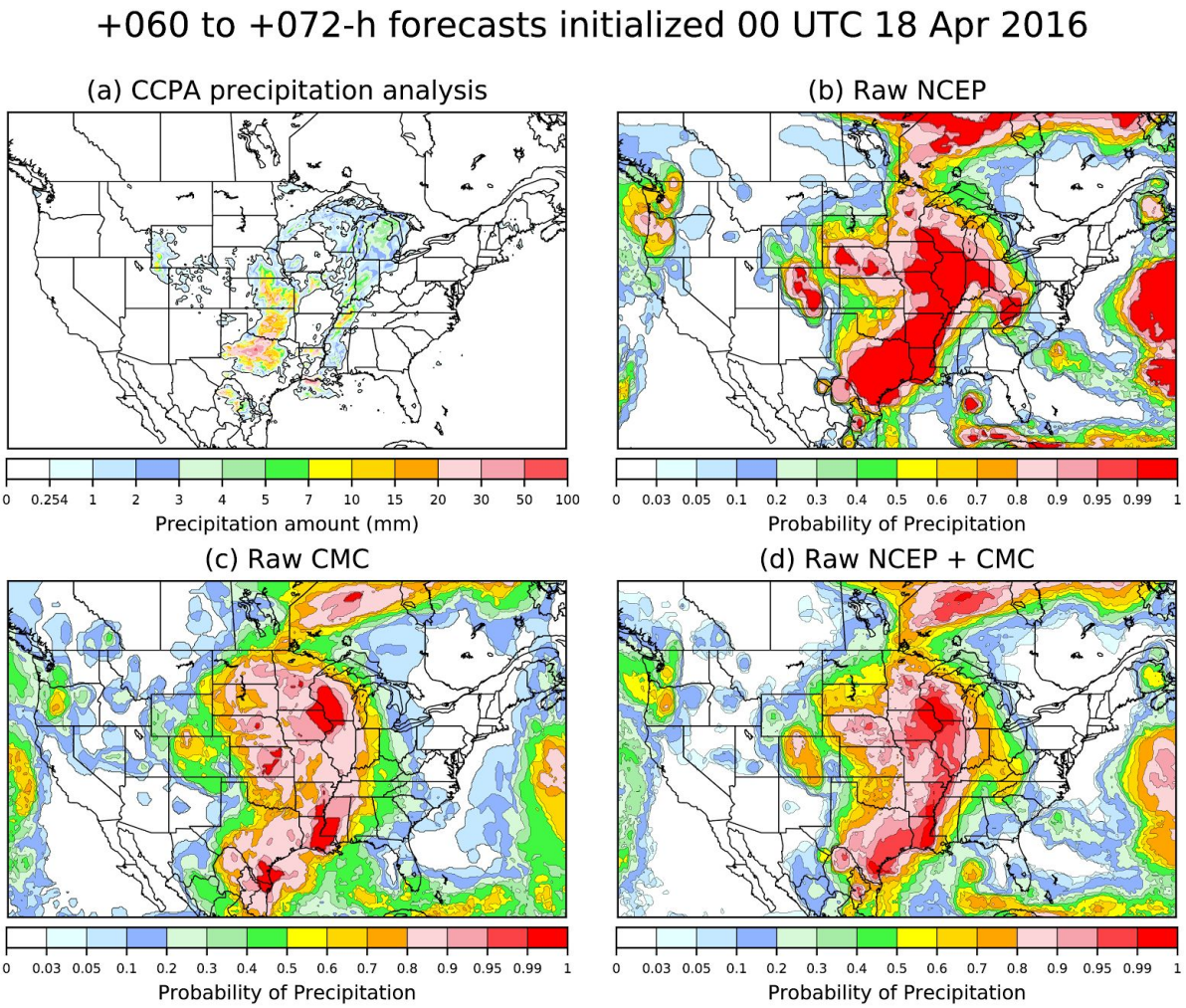

Figure 7.10. Case study of the steps in POP12 postprocessing for a +60 to +72 h forecast initialized at 00 UTC 18 April 2016. (a) CCPA precipitation analysis. (b) Raw NCEP POP12 forecast. (c) Raw CMC POP12 forecast. (d) Raw CMC+NCEP POP12 forecast.

Note to copy editor: this figure and Fig. 7.11 below would be improved if rotated by 90 degrees and plotted to fill up a whole page, thereby increasing the size.

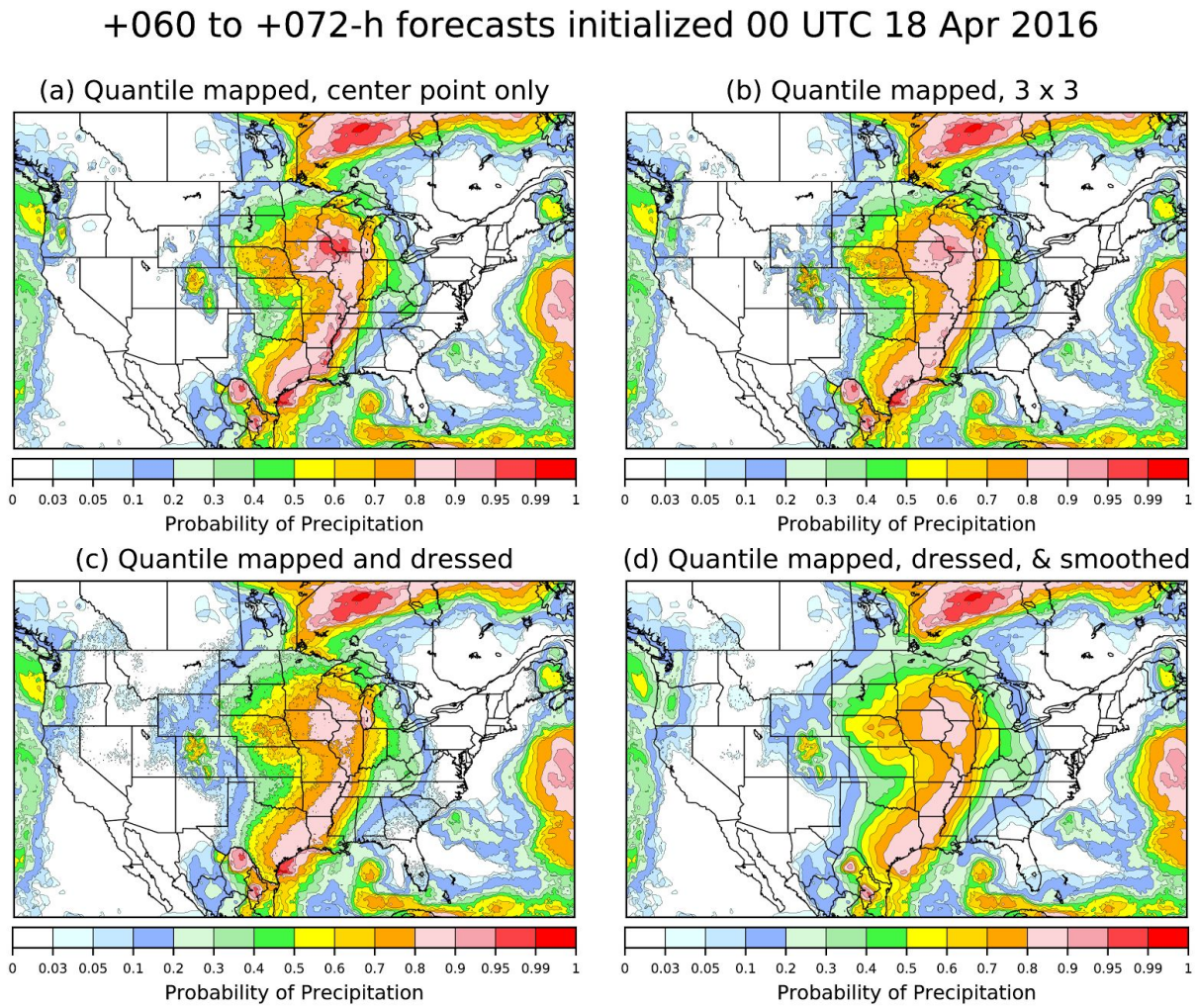

Figure 7.11. Continuation of the case study of the steps in POP12 postprocessing for a +60 to +72 h forecast initialized at 00 UTC 18 April 2016. (a) postprocessing with quantile mapping using only the grid point in question. (b) Quantile mapping of a 3x3 array of grid points, centered on each point of interest. (c) Quantile-mapped and dressed POP12 forecast, and (d) the final product, with 3x3 quantile mapping, dressing, and smoothing.

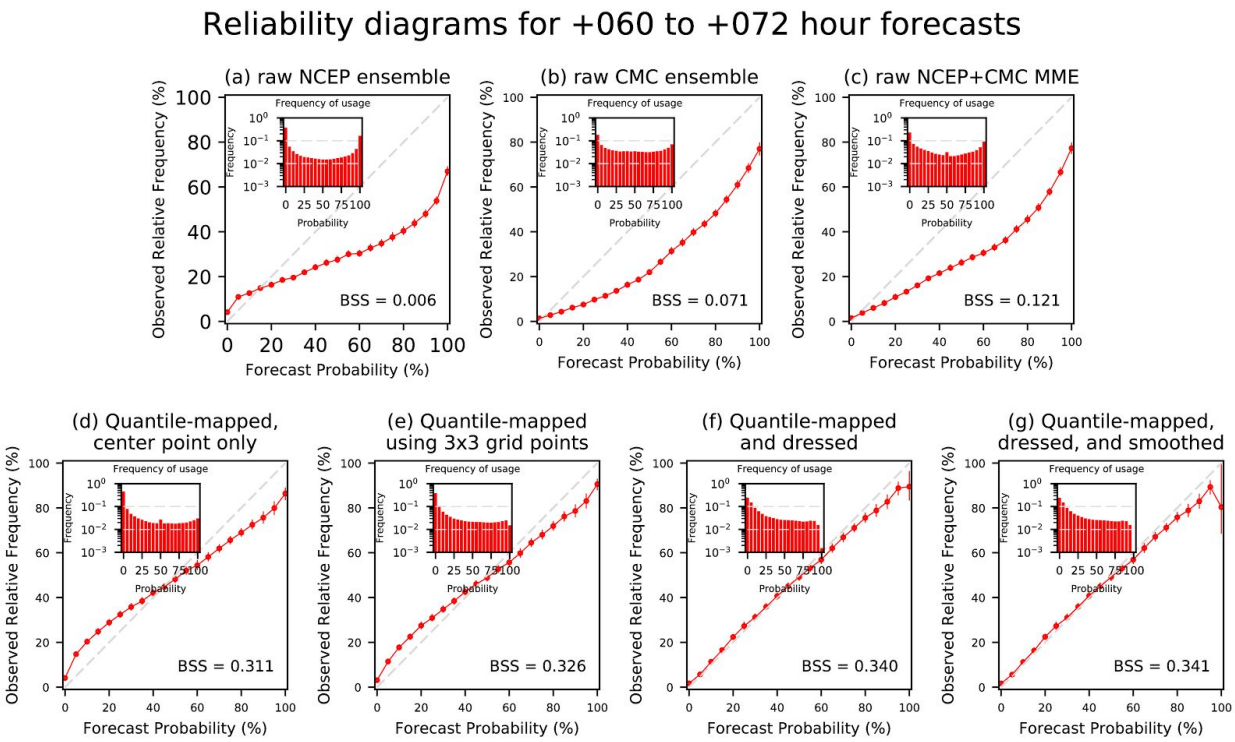

Figure 7.12: Reliability diagrams for +60 to +72 hour POP12 forecasts over the CONUS. Inset histograms show overall frequency with which forecasts are issued, and Brier Skill Scores are noted. (a) Raw NCEP ensemble forecasts, (b) raw CMC ensemble forecasts, (c) raw multi-model ensemble forecasts, (d) post-processed guidance after stochastic quantile mapping using the center point only, (e) after stochastic quantile mapping using 3 × 3 stencil of points, (f) after dressing, and (g) after smoothing. Error bars represent the 5th and 95th percentiles of from a 1000- sample bootstrap distribution generated by sampling case days with replacement.

7.6 Collaborating on software and test data to accelerate postprocessing improvement.

Finally, let us turn attention to possible ways that the statistical post-processing community can work more effectively together than individually. Recognizing the complexity of developing weather prediction components, data assimilation systems and forecast models are increasingly maintained and supported as community endeavors (Skamarock et al. 2008, COSMO 2016). Users are free to download the code, compile it on the computer system of their choice, generate assimilations and forecasts, and develop and test algorithmic improvements. Commonly with such systems, there is a protocol for submitting algorithmic changes to be incorporated back into the community software. If they are coded to predefined standards and demonstrated to improve the forecasts, a change review board can accept these software modifications, which are then incorporated into future releases.

Envision a similar community infrastructure for postprocessing, including a software and test-data repository. Code for reading these data and writing post-processed output would be available. A variety of post-processing algorithms, verification routines, and data visualization

tools would be part of the software library. Data sets for common problems of interest would be available in portable formats such as netCDF (Unidata, 2012).

With these components in place, answering research questions could become much more straightforward. A university investigator that seeks to develop a new post-processing methodology and to test their hypothesis that the method improves upon existing methods could start with data sets and a code infrastructure in place. Their time and effort could be concentrated on the science question at hand, not wrangling with the data and supporting code. Comparisons against existing benchmarks would be straightforward, and presuming confirmation of the hypothesis of an improved method, the resulting journal articles would be more valuable. Readers would have confidence that the new method had been sufficiently demonstrated to provide an improvement over existing standards.

How do we go about building such a community? This topic was discussed at a 2016 workshop on postprocessing, hosted by the US NWS; the workshop recommendations shown in the sidebar (copy editor: take material from the Appendix for this sidebar). They require some moderate amount of resources and commitment from a few key individuals, hopefully supported by one or more weather prediction organizations. With the rapid growth of open-source software, there are many established best practices that can be followed to ensure that our new community would have a greater likelihood of flourishing. Standard software version control systems such as “git” would be used; these allow a new user to replicate (create a branch of) the community software on their local computer system, make modifications, but never lose the original. A governance procedure would be established to enable diverse groups to work together and make decisions about software and data changes. Following established best practices, a ticket-tracking system would be established to monitor suggested product improvements and their disposition. A change-control board would be instituted to manage code contributions. These code contributions would be expected to follow predefined testing, documentation, and metadata standards. Documentation would be centralized and consolidated into a few core documents. Ideally, assistance would be provided to help collaborators.

How might a diverse suite of software be maintained such that it serves the joint needs of a more free-wheeling academic community and the more controlled needs of weather services? How do we build a community that fosters intellectual diversity while containing software entropy, an uncontrolled growth of code size and diversity? Following a suggestion by Tom Auligné (personal communication, 2016), a software repository might have several tiers. Users could contribute their modified software branch back to an outer tier of the repository, home for a wide diversity of software branches that could be shared between investigators working on a common problem. Software in this tier would not be rigorously vetted. Should a community user wish to see the software become incorporated into the community “trunk,” then the software would be reviewed by a change-control board that evaluates the software for coding clarity, adherence to documentation and test standards, and results. Presuming acceptance, the software would then become part of a more limited suite of broadly supported

community algorithms. A final inner tier would be for a particular agency like the US NWS. The software that is run on operational supercomputers would be subject to further refinement and quality control, per standards established by that agency. Conceivably, one might envision multiple agencies having similar or slightly different inner tiers, but sharing community contributions migrating in from the middle tier.

7.7. Recommendations and conclusions.

This chapter discussed many of the practical aspects related to statistical postprocessing. To achieve the highest quality result, the statistician must attend to the practical realities of the data to be used in addition to the algorithmic design of the post-processing software. Here are some recommendations on how we can make more rapid progress:

(a) *Post-processing scientists should engage with prediction system developers about the data needs*. There can be tension between the desires of our model-development colleagues to improve the prediction system as quickly as possible and the desire of statisticians for guidance that are homogeneous in their error characteristics as well as high in quality. Finding the balance is challenging. The earlier the data requirements are communicated with the prediction system developers, the easier it will be for them to accommodate post-processing needs in their plans. For example, the prediction system developers may have a choice between two possible upgrade paths, one that minimizes RMS error at the expense of some bias, or one that minimizes forecast bias at the expense of slightly higher RMS error. Changes in bias from one model version to the next are generally more challenging to address in postprocessing. Hence, if we agree that the end goal is high-quality post-processed guidance rather than the lowest-error raw guidance, then the latter upgrade path (minimizing forecast bias) may be a more sensible path forward.

(b) *Challenge ourselves to use the existing training data more efficiently*. Given the expense and work required to generate lengthy retrospective data, our algorithms should extract the most information possible from the limited data at hand.

(c) *Work together to build a postprocessing community, sharing data and software*. By building algorithms in isolation, we are unsure as to whether our design represents an improvement over existing methods. If standard data sets are available to researchers, and if we share code for data input, output, and verification, then it becomes much simpler to test our methodology against other reference standards. Rome wasn't built in a day; this is an ambitious goal, but we can start with simple and productive steps. After we finish a project, we can make our data freely available and share our algorithms in public portals such as github. Here is my personal example, a reforecast precipitation data set and associated analog method software (https://github.com/ThomasMoreHamill/analog) .

(d) More generally, *prediction centers should share their data*. TIGGE (Bougeault et al. 2009, Swinbank et al. 2016) was an international research project that archived global

ensemble data for research purposes. Many scientists have used data from the TIGGE archive to demonstrate the improvements possible from the postprocessing of multi-center ensemble data. We all have much to gain and little to lose by sharing more data in real time, leveraging each others' investments in research and computing. In particular, such data is of exceptional importance to developing countries that cannot yet afford to develop their own prediction systems.

(e) *Collaborate with professional statisticians*. Chances are you, the reader, are trained as an atmospheric scientist or hydrologist. You have knowledge about data characteristics and potential predictors that a statistician will not have. However, a statistician is likely to have a more thorough grounding in Bayesian methods, in spatial statistics, in machine learning. Together you may be able to produce higher-quality products than you could working individually.

Appendix: Recommendations from workshop on statistical postprocessing.

Here are some general recommendations from the February 2016 workshop on statistical postprocessing (http://www.dtcenter.org/events/workshops16/post-processing/). Many focus on the actions that government weather prediction centers should take to support the post-processing community. Recommendations focus on science, community infrastructure, and data.

Science:

1. Entrain professional statisticians to assist meteorologists with the development of statistical improved post-processing methodologies.
2. Perform more intercomparisons of existing algorithms to determine which are the most skillful and reliable.
3. Standard-setting algorithms should be coded so that they are efficient and easily usable by the broader community.
4. Given the challenges with developing and storing high-quality training data sets, further research and development is particularly needed on methodologies that permit high-quality results to be developed with minimal training data.
5. Algorithms developed in the future should be validated against relevant standards of comparison such as those developed in (2) - (3) above.

Community infrastructure.

The postprocessing community should collaborate to build high-quality shared code and data repositories and should maintain this. Ideally, the community repository would have characteristics such as:

- Tracking of software changes using a community-standard version control system such as git.
- Support for tiers in the repository, from tightly controlled (operational prediction code) to more loosely controlled (community scientists).
- An established process to manage incorporation code contributions into the inner tiers of the repository, i.e., a change control board.
- A ticket-tracking system to monitor requested code changes and their disposition.
- Established standards for metadata, tests, and documentation.
- Use of an agreed-upon common vocabulary.
- A centralized location for documentation and data access, with focus on a few core documents.
- Use two or three modern common data formats that should be used (e.g., netCDF, HDF, geoJSON) that will satisfy operational, research, collaboration, and archival purposes.

Data:

1. Prediction centers, if possible, should regularly generate high-quality reanalysis and reforecast data.

2. Prediction centers should ensure that future high-performance computing and disk procurement reflect the compute and storage needs reforecasts and reanalyses.

3. Prediction centers should also generate high-quality, high-resolution analysis data for training and validation.

4. Prediction centers should post-process and make readily available commonly used “foundational data” (e.g., temperature, precipitation) for use insider the weather service and across the broader enterprise.

5. Prediction centers should make training data easily accessible.

6. Given the challenges with transmission of voluminous training data, prediction centers are encouraged to either set aside computational resources for external collaborators working on postprocessing, or to permit read access to storage systems with training data.

7. Survey of post-processing product developers to ensure prediction centers are saving on disk the relevant predictor information.

References

Allen, R. L., and M.C. Erickson, 2001: AVN-based MOS precipitation type guidance for the United States. *NWS Technical Procedures Bulletin* No. 476, NOAA, U.S. Dept. of Commerce, 9pp.

Anthes, R., D. Ector, D. Hunt, Y. Kuo, C. Rocken, W. Schreiner, S. Sokolovskiy, S. Syndergaard, T. Wee, Z. Zeng, P. Bernhardt, K. Dymond, Y. Chen, H. Liu, K. Manning, W. Randel, K. Trenberth, L. Cucurull, S. Healy, S. Ho, C. McCormick, T. Meehan, D. Thompson, and N. Yen, 2008: The COSMIC/FORMOSAT-3 mission: early results. *Bull. Amer. Meteor. Soc.,* **89**, 313–333, doi: 10.1175/BAMS-89-3-313.

Auligné, T., McNally, A. P. and Dee, D. P., 2007: Adaptive bias correction for satellite data in a numerical weather prediction system. *Quart. J. Royal Meteor. Soc*., **133**, 631–642. doi:10.1002/qj.56

Benjamin, S.G., B. D. Jamison, W. R. Moninger, S. R. Sahm, B. E. Schwartz, and T. W. Schlatter, 2010: Relative short-range forecast impact from aircraft, profiler, radiosonde, VAD, GPS-PW, METAR, and mesonet observations via the RUC hourly assimilation cycle. *Mon. Wea. Rev.,* **138**, 1319–1343, doi: 10.1175/2009MWR3097.1.

Bi, L., J. A. Jung, M. C. Morgan, and J. F. Le Marshall, 2011: Assessment of assimilating ASCAT surface wind retrievals in the NCEP Global Data Assimilation System. *Mon. Wea. Rev.,***139,** 3405–3421.

Boisserie, M., Decharme, B., Descamps, L. and Arbogast, P., 2016: Land surface initialization strategy for a global reforecast dataset. *Quart. J. Royal Meteor. Soc*., **142**, 880–888. doi:10.1002/qj.2688 .

Bougeault, P., and others, 2009: TheTHORPEX Interactive Grand Global Ensemble (TIGGE). *Bull. Amer. Meteor. Soc*., **91**, 1059-1072.

Buehner, M., J. Morneau, and C. Charette, 2013: Four-dimensional ensemble–variational data assimilation for global deterministic weather prediction. *Nonlinear Processes Geophys.*, **20**, 669–682, doi:10.5194/npg-20-669-2013.

Candille, G., 2009: The Multiensemble Approach: The NAEFS Example. *Mon. Wea. Rev.,* **137**, 1655–1665, https://doi.org/10.1175/2008MWR2682.1 .

Collard, A. D. and McNally, A. P. , 2009: The assimilation of infrared atmospheric sounding interferometer radiances at ECMWF. *Quart. J. Royal Meteor. Soc*., **135**, 1044–1058. doi:10.1002/qj.410

Courtier, P., Thépaut, J.-N. and Hollingsworth, A., 1994: A strategy for operational implementation of 4D-Var, using an incremental approach. *Quart. J.Royal Meteor. Soc.*, **120**, 1367–1387. doi:10.1002/qj.49712051912.

Cui, B., Z. Toth, Y. Zhu, and D. Hou, 2012: Bias correction for global ensemble forecast. *Wea. Forecasting,* **27**, 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. *Quart. J. Royal Meteor. Soc.*, to appear. DOI: 10.1002/qj.2975

Daley, R., 1991: *Atmospheric Data Analysis*. Cambridge University Press. 457 pp.

Dee, D. P., 2005: Bias and data assimilation. *Quart. J. Royal Meteor. Soc.*, **131**, 3323–3343. doi:10.1256/qj.05.137

Dee, D. P., and coauthors, 2011: The ERA-Interim reanalysis: configuration and performance of the data assimilation system. *Quart. J. Royal Meteor. Soc.*, **137**, 553–597. doi:10.1002/qj.828

Dee., D. P., M. Balmaseda, G. Balsamo, R. Engelen, J. Simmons, and J.-N. Thépaut, 2014: Towards a consistent reanalysis of the climate system. *Bull. Amer. Meteor. Soc*., 95, 1236-1248. DOI:10.1175/BAMS-D-13-00043.1

De Pondeca, M., G. Manikin, G. DiMego, S. Benjamin, D. Parrish, R. Purser, W. Wu, J. Horel, D. Myrick, Y. Lin, R. Aune, D. Keyser, B. Colman, G. Mann, and J. Vavra, 2011: The real-time mesoscale analysis at NOAA's National Centers for Environmental Prediction: current status and development. *Wea. Forecasting,* **26**, 593–612, doi: 10.1175/WAF-D-10-05037.1.

Durran, D. R., 2010: *Numerical Methods for Fluid Dynamics (2nd Ed.)*. Springer, 516 pp.

Evensen, G., 2014: Data Assimilation, The Ensemble Kalman Filter (2nd Edition). Springer, 307 pp. **ISBN-13:** 978-3642424762.

Fortin, V., Favre, A.-c. and Saïd, M., 2006: Probabilistic forecasting from ensemble prediction systems: Improving upon the best-member method by using a different weight and dressing kernel for each member. *Quart.J. Royal Meteor. Soc*., **132**, 1349–1369. doi:10.1256/qj.05.167

Gagnon, N., X. Deng, P.L. Houtekamer, S. Beauregard, A. Erfani, M. Charron, R. Lahlo, and J. Marcoux 2014: *Improvements to the Global Ensemble Prediction System (GEPS) from version 3.1.0 to version 4.0.0.* Environment Canada Tech Note, available at http://collaboration.cmc.ec.gc.ca/cmc/cmoi/product_guide/docs/lib/technote_geps-400_20141118_e.pdf .

——, X. Deng, P. L. Houtekamer, A. Erfani, M. Charron, S. Beauregard, R. Frenette, D. Racette, and R. Lahlou, 2015: *Improvements to the Global Ensemble Prediction System from version 4.0.1 to version 4.1.1.* Environment Canada Tech Note, available at http://collaboration.cmc.ec.gc.ca/cmc/cmoi/product_guide/docs/lib/technote_geps-411_20151215_e.pdf .

Glahn, B., K. Gilbert, R. Cosgrove, D. Ruth, and K. Sheets, 2009: The Gridding of MOS. *Wea. Forecasting,* **24**, 520–529, doi: 10.1175/2008WAF2007080.1

Hagedorn, R., Buizza, R., Hamill, T. M., Leutbecher, M., and T. N. Palmer, 2012: Comparing TIGGE multi-model forecasts with reforecast-calibrated ECMWF ensemble forecasts. *Quart J. Royal Meteor Soc.*, **138**, 1814-1827.

Hamill, T. M., J. S. Whitaker, and X. Wei, 2004: Ensemble re-forecasting: improving medium-range forecast skill using retrospective forecasts. *Mon. Wea. Rev.*, **132**, 1434-1447.

Hamill, T. M., J. S. Whitaker, and S. L. Mullen, 2006: Reforecasts, an important dataset for improving weather predictions. *Bull. Amer. Meteor. Soc.*, **87**,33-46.

Hamill, T. M., 2006: Ensemble-based atmospheric data assimilation. Chapter 6 of *Predictability of Weather and Climate*, Cambridge Press, 124-156.

Hamill, T. M., 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 data set. *Bull Amer. Meteor. Soc.*, **94**, 1553-1565.

Hamill, T. M., 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. Also: online appendix A and appendix B.

Hamill, T. M., E. Engle, D. Myrick, M. Peroutka, C. Finan, and M. Scheuerer, 2017: The US National Blend of Models statistical post-processing of probability of precipitation and deterministic precipitation amount. *Mon. Wea. Rev.*, conditionally accepted. Also: online Appendix A and online Appendix B.

Hamill, T. M., 2017: Changes in the systematic errors of global reforecasts due to an evolving data assimilation system. *Mon. Wea. Rev.*, submitted. Available from tom.hamill@noaa.gov.

Hastie, T. J., and R. J. Tibshirani, 1990: *Generalized Additive Models*. Chapman and Hall, 335 pp.

Hastie, T., R. Tibshirani, and J. Friedman, 2001: *The Elements of Statistical Learning*. Springer, 533 pp.

Hodyss, D., E. Satterfield, J. McLay, T. Hamill, and M. Scheuerer, 2016: Inaccuracies with Multimodel Postprocessing Methods Involving Weighted, Regression-Corrected Forecasts. *Mon. Wea. Rev.,* **144**, 1649–1668, doi: 10.1175/MWR-D-15-0204.1.

Hopson, T. M., and P. J. Webster, 2010: A 1–10-day ensemble forecasting scheme for the major river basins of Bangladesh: Forecasting severe floods of 2003–07. *J. Hydrometeor.*, **11**, 618–641, doi:10.1175/2009JHM1006.1.

Hou, D., M. Charles, Y. Luo, Z. Toth, Y. Zhu, R. Krzysztofowicz, Y. Lin, P. Xie, D. Seo, M. Pena, and B. Cui, 2014: Climatology-Calibrated Precipitation Analysis at Fine Scales: Statistical Adjustment of Stage IV toward CPC Gauge-Based Analysis. *J. Hydrometeor.,* **15**, 2542–2557, doi: 10.1175/JHM-D-11-0140.1.

Kalnay, E., 2003: *Atmospheric Modeling, Data Assimilation, and Predictability*. Cambridge University Press, 341 pp.

Kleiber, W., A. E. Raftery, J. Baars, T. Gneiting, C. F. Mass, and E. Grimit, 2011: Locally calibrated probabilistic temperature forecasting using geostatistical model averaging and local Bayesian model averaging. *Mon. Wea. Rev*., **139**, 2630-2649.

Kleiber, W., Raftery, A.E. and Gneiting, T., 2011: Geostatistical model averaging for locally calibrated probabilistic quantitative precipitation forecasting. *J. Amer. Stat. Assoc*., **106**, 1291-1303.

Kleist, D. T., and K. Ide, 2015: An OSSE-based evaluation of hybrid variational–ensemble data assimilation for the NCEP GFS. Part I: System description and 3D-hybrid results. *Mon. Wea. Rev.*, **143**, 433–451, doi:10.1175/MWR-D-13-00351.1.

Lalaurette, F., 2003: Early detection of abnormal weather conditions using a probabilistic extreme forecast index. *Quart. J.Royal Meteor. Soc.*, **129**, 3037–3057. doi:10.1256/qj.02.152

Lerch, S., and S. Baran, 2017: Similarity-based semilocal estimation of post-processing models. *J. Royal Stat. Soc. C*, **66,** 29-51. doi:10.1111/rssc.12153

Lespinas, F., V. Fortin, G. Roy, P. Rasmussen, and T. Stadnyk, 2015: Performance Evaluation of the Canadian Precipitation Analysis (CaPA). *J. Hydrometeor.,* **16**, 2045–2064, doi: 10.1175/JHM-D-14-0191.1.

Lin, H., N. Gagnon, S. Beauredard, R. Muncaster, M. Markovic, B. Denis, and M. Charron, 2016: GEPS based monthly prediction at the Canadian Meteorological Centre. *Mon. Wea. Rev*., in press. DOI: http://dx.doi.org/10.1175/MWR-D-16-0138.1

Maraun, D., 2013: Bias correction, quantile mapping, and downscaling: revisiting the inflation issue. *J. Climate*, **26**, 2137-2143. http://dx.doi.org/10.1175/JCLI-D-12-00821.1

Mass, C. F., J. Baars, G. Wedam, E. Grimit, and R. Steed, 2008: Removal of systematic model bias on a model grid. *Wea. Forecasting*, **23**, 438-459.

McNally, A. P., Watts, P. D., A. Smith, J., Engelen, R., Kelly, G. A., Thépaut, J. N. and Matricardi, M., 2006: The assimilation of AIRS radiance data at ECMWF. Quart. J. Royal Meteor. Soc., **132**, 935–957.

NCAR/MMM, 2017: ARW Version 3 Modeling System User's Guide. 434 pp. Available at http://www2.mmm.ucar.edu/wrf/users/pub-doc.html ,

Neter, J., W. Wasserman, and M.H. Kutner, 1990: *Applied Linear Statistical Models (3rd Edition)*. Irwin Press, 1181 pp.

Ou, M., M. Charles, and D. Collins, 2016: Sensitivity of Calibrated Week-2 Probabilistic Forecast Skill to Reforecast Sampling of the NCEP Global Ensemble Forecast System. *Wea. Forecasting,* **31**, 1093–1107, doi: 10.1175/WAF-D-15-0166.1.

Park, Y.-Y., R. Buizza, and M. Leutbecher, 2008: TIGGE: Preliminary results on comparing and combining ensembles. *Quart. J. Roy. Meteor. Soc.*, 134, 2029– 2050, doi:10.1002/qj.334.

Petroliagis, T. I., and P. Pinson, 2014: Early warnings of extreme winds using the ECMWF Extreme Forecast Index. *Met. Apps*., **21,** 171–185. doi:10.1002/met.1339

Press, W. H., S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, 1992: *Numerical Recipes in Fortran (2nd Ed)*. Cambridge Press, 963 pp.

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.

Richardson, D. S., 2001: Measures of skill and value of ensemble prediction systems, their interrelationship and the effect of ensemble size. *Quart. J. Royal Meteor. Soc.*, **127**, 2473-2489.

Saha, S., and Coauthors, 2010: The NCEP Climate Forecast System Reanalysis. *Bull. Amer. Meteor. Soc*., **91**, 1015–1057.

Schaake, J. C., T. M. Hamill, R. Buizza, and M. Clark, 2007: HEPEX, the Hydrological Ensemble Prediction Experiment. *Bull. Amer. Meteor. Soc.*, **88**, 1541-1547.

Schättler, U., G. Doms, and C. Schraff, 2016: *A Description of the Nonhydrostatic Regional COSMO-Model. Part VII: Users Guide*. 221 pp. Available at http://www2.cosmo-model.org/content/model/documentation/core/ .

Scheuerer, M. and Büermann, L., 2014: Spatially adaptive post-processing of ensemble forecasts for temperature. J. Royal. Stat. Soc. C, **63**, 405–422. doi:10.1111/rssc.12040

Scheuerer, M. and G. König, 2014: Gridded, locally calibrated, probabilistic temperature forecasts based on ensemble model output statistics. *Quart. J. Royal Meteor. Soc*., **140**, 2582-2590.

Scheuerer, M. and D. Möller, 2015: Probabilistic wind speed forecasting on a grid based on ensemble model output statistics. *The Annals of Applied Statistics*, **9(3)**, 1328-1349.

Scheuerer, M., and T. M. Hamill, 2015: Statistical post-processing of ensemble precipitation forecasts by fitting censored, shifted Gamma distributions. *Mon. Wea. Rev*., **143**, 4578-4596.

Skamarock, W. C., J. B. Klemp, J. Dudhia, D. O. Gill, D. M. Barker, W. Wang, J. G. Powers, 2005: A Description of the Advanced Research WRF Version 2. NCAR TECHNICAL NOTE, NCAR/TN–468+STR, 88 pp. Available from http://www2.mmm.ucar.edu/wrf/users/pub-doc.html .

Sloughter, J. M., A. E. Raftery, T. Gneiting, and C. Fraley, 2007: Probabilistic quantitative precipitation forecasting using Bayesian Model Averaging. *Mon. Wea. Rev*., **135**, 3209-3220. DOI: http://dx.doi.org/10.1175/MWR3441.1

Stauffer, R., N. Umlauf, J. W. Messner, G. J. Mayr, and A. Zeileis, 2017: Ensemble postprocessing of daily precipitation sums over complex terrain using censored high-resolution standardized anomalies. *Mon. Wea. Rev*., **145**, 955-969. DOI: 10.1175/MWR-D-16-0260.1

Stensrud, D. J., 2007: *Parameterization Schemes. Keys to Understanding Numerical Weather Prediction Models.* Cambridge Press, 459 pp.

Swinbank, R., and others, 2016: The TIGGE project and its achievements. *Bull. Amer. Meteor. Soc.*, **97**, 49-67, doi: 10.1175/BAMS-D-13-00191.1.

Unidata, 2012: Integrated Data Viewer (IDV) version 3.1 [software]. Boulder, CO: UCAR/Unidata. (http://doi.org/10.5065/D6RN35XM and https://www.unidata.ucar.edu/software/netcdf/)

Uppala, S. M., KÅllberg, P. W., Simmons, A. J., Andrae, U., Bechtold, V. D. C., Fiorino, M., Gibson, J. K., Haseler, J., Hernandez, A., Kelly, G. A., Li, X., Onogi, K., Saarinen, S., Sokka, N., Allan, R. P., Andersson, E., Arpe, K., Balmaseda, M. A., Beljaars, A. C. M., Berg, L. V. D.,

Bidlot, J., Bormann, N., Caires, S., Chevallier, F., Dethof, A., Dragosavac, M., Fisher, M., Fuentes, M., Hagemann, S., Hólm, E., Hoskins, B. J., Isaksen, L., Janssen, P. A. E. M., Jenne, R., McNally, A. P., Mahfouf, J.-F., Morcrette, J.-J., Rayner, N. A., Saunders, R. W., Simon, P., Sterl, A., Trenberth, K. E., Untch, A., Vasiljevic, D., Viterbo, P. and Woollen, J. (2005), The ERA-40 re-analysis. *Quart. J. Royal Meteor. Soc.*, **131**, 2961–3012. doi:10.1256/qj.04.176

Vannitsem, S., and R. Hagedorn, 2011: Ensemble forecast post-processing over Belgium: Comparison of deterministic-like and ensemble regression methods. *Meteorol. Appl.*, **18**, 94-104.

Velden, C. S., J. Daniels, D. Stettner, D. Santek, J. Key, J. Dunion, K. Holmlund, G. Dengel, W. Bresky, and P. Menzel, 2005: Recent nnovations in deriving tropospheric winds from meteorological satellites. *Bull. Amer. Meteor. Soc.*, **86**, 205-223.

Vitart. F., G. Balsamo, R. Buizza, L. Ferranti, S. Keeley, L. Magnusson, F. Molteni and A. Weisheimer, 2014: *Sub-seasonal predictions*. ECMWF Tech Memo 738. Available at: https://www.ecmwf.int/sites/default/files/elibrary/2014/12943-sub-seasonal-predictions.pdf

Voisin, N., J. C. Schaake, and D. P. Lettenmaier, 2010: Calibration and downscaling methods for quantitative ensemble precipitation forecasts. *Wea. Forecasting,* **25**, 1603-1627.

Warner, T. T., 2011: *Numerical Weather and Climate Prediction*. Cambridge University Press, 526 pp.

Wikipedia, 2016: (https://en.wikipedia.org/wiki/Bias%E2%80%93variance_tradeoff)

Wilks, D. S., 2006: Comparison of ensemble-MOS methods in the Lorenz '96 setting. *Meteor. Apps.*, **13**, 246-256.

Wilks, D. S., 2011: *Statistical Methods in the Atmospheric Sciences (3rd Ed.)*. Academic Press, 676 pp.

Zhang, J., and coauthors, 2016: Multi-Radar Multi-Sensor (MRMS) Quantitative Precipitation Estimation: Initial Operating Capabilities. *Bull. Amer. Meteor. Soc.,* **97**, 621–638, doi: 10.1175/BAMS-D-14-00174.1.

Zhou, X., Y. Zhu, Y. Luo, J. Peng and R. Wobus, 2016: The NCEP Global Ensemble Forecast System with EnKF. *Mon. Wea. Rev.*, submitted. Available from xiaqiong.zhou@noaa.gov.