13 Statistical post processing of model output

13.1 Background

The statistical post processing, or calibration, of operational NWP-model output is common because it can result in skill metrics that are equivalent to many years of improvement to the basic model. And, the greater skill is achieved at relatively little day-to-day expense, compared to other traditional approaches of trying to improve skill, such as through increasing the model resolution.

Historically, statistical post-processing methods were used to diagnose variables that could not be predicted directly by the low-resolution, early-generation NWP models. Standard model dependent variables associated with the large-scale conditions were statistically related to other poorly predicted or unpredicted weather variables such as freezing rain, fog, and cloud cover. However, many current-generation, high-resolution models can explicitly forecast such variables, and statistical correction methods are primarily employed to reduce systematic errors.

There is a variety of ways of classifying statistical post-processing methods. They may be categorized in terms of the statistical techniques used, as well as by the types of predictor data that are used for development of the statistical relationships. And, distinctions are made between static and dynamic methods. With static methods, statistical algorithms are developed for removing systematic error using a long training period that is based on the same version of the model, and the algorithms are applied without change for a significant period of time. Because of the computational expense associated with the calculation of the statistical relationships, models cannot be upgraded frequently because doing so requires recalculation of the relationships. Even when significant code errors are revealed, they cannot be corrected until new relationships are created. In contrast, with dynamic methods the calibration equations are recalculated on a regular basis.

Statistical post-processing methods are not appropriate for use in research applications of models. For physical-process studies, it is important that the model output be consistent with the dynamic equations, so artificial adjustments in the output would not be appropriate. Also, in such studies it is straightforward to optimize the model for a particular case to reduce systematic error (e.g., by testing different physical-process parameterizations, model resolutions, etc.), so there is less need for statistical adjustment than with operational NWP. And, research is often aimed at improving the model in order to reduce the systematic error.

Downloaded from https://www.cambridge.org/core. The University of British Columbia Library, on 14 Jan 2020 at 06:11:04, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511763243.014

The NWP-model forecast variables are sometimes used as input to specialized models that provide information about other quantities. For example, air-quality models contain continuity equations for various gaseous and aerosol species, they calculate the transport and diffusion of these contaminants, and they represent their chemical transformations. This use of appended and specialized models may be viewed as one type of post processing of the NWP-model output, but it is sufficiently specialized that it will be treated separately in Chapter 14.

The following section reviews some different approaches for statistical correction of model output. Another type of model post processing employs what are called *weather generators*, which are summarized in Section 13.3. Weather generators, also discussed in Chapter 16 in the context of climate models, take model output that is typically smoother than reality in terms of space and time variability, and define a more-realistic statistical structure. This kind of post processing is important for some model applications, for example in predicting flooding where the local short-time-scale variations in rainfall intensity are highly relevant for estimating the partitioning of rainfall between runoff and infiltration. In the last section is a brief discussion of how some types of downscalings of model output – that is, processing that can define the modulation of the large scales by local forcing, such as from orography – represent a form of statistical post processing.

13.2 Systematic-error removal

The following subsections review various methods for statistically correcting NWP-model forecasts in order to reduce the systematic error. The static methods require the use of a lengthy period of model reforecasts in order to define the statistical corrections based on relationships between past model output and past observations. These statistical relationships are not updated frequently. In contrast, the dynamic methods perform corrections to forecasts based on much shorter periods of training. In both cases, the goal is to reduce the error in current forecasts using estimates of past error.

Note that only systematic error is reduced by these methods. The random errors that result from numerically induced phase errors in the propagation of features, the smoothing of small-scale propagating features associated with insufficient model resolution, and other causes, will remain in the solution and cannot be statistically removed. However, the systematic error can represent a significant fraction of the total error, especially near the ground, so removing it through the use of post-processing methods can be very beneficial. For example, Fig. 13.1 shows the systematic and random forecast errors in the near-surface temperature (2 m AGL) and wind speed (10 m AGL), based on regional mesoscale-model simulations for the southwestern USA. The systematic temperature errors are clearly larger at some of the observation sites, presumably because local forcing is not represented well in the model.

Downloaded from https://www.cambridge.org/core. The University of British Columbia Library, on 14 Jan 2020 at 06:11:04, subject to the Cambridge Core terms of use , available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511763243.014

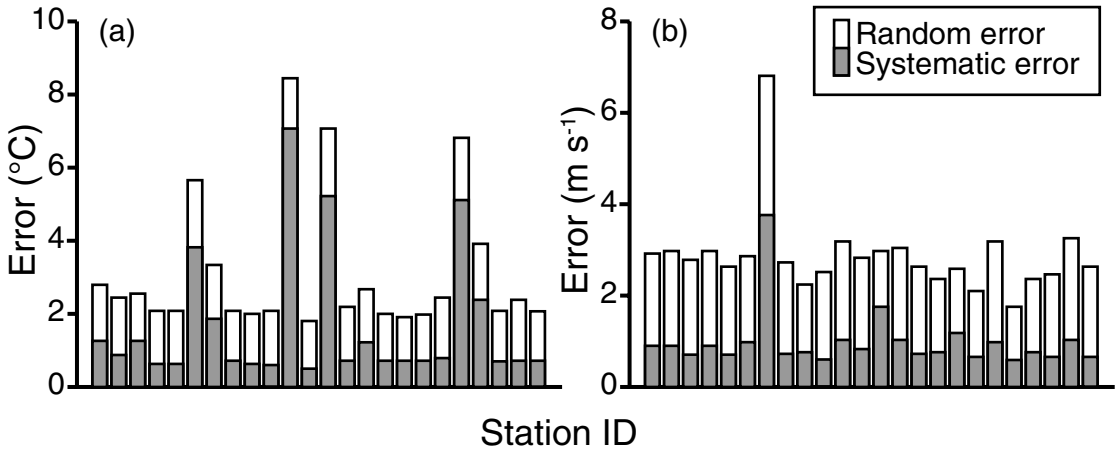

Fig. 13.1 Systematic and random errors in the near-surface temperature (a, 2 m AGL) and wind speed (b, 10 m AGL) calculated based on regional mesoscale-model simulations for southwestern USA. Each bar corresponds to a different observation location. Adapted from Hacker and Rife (2007).

13.2.1 The "perfect-prog" method

The earliest approach to statistical post processing is known as the Perfect-Prog (PP, perfect-prognosis) method (Klein *et al.* 1959). Here, observations of quantities that are predicted by the model (the predictors) are statistically related to observations of a predictand that may or may not be predicted by the model. The regression relationships are then applied to NWP-model forecasts of the predictors to produce forecasts of the predictands. Because the statistical relationships are not generated using model forecasts, they do not correct for model error. They simply statistically translate predicted variables into unpredicted or poorly predicted variables. In effect, it is assumed that the model prognosis is perfect. Because, as noted earlier, current models can explicitly predict many of the quantities that previously had to be statistically inferred through the PP method, this approach is less used operationally. A benefit of the PP method is that it is not dependent on the model to which it is applied, and thus the statistical relationships do not need to be recalculated when the model is modified. Figure 13.2 schematically compares the PP approach with the method of Model Output Statistics (MOS) described in the next section.

13.2.2 Model output statistics

The calculation of MOS involves statistically relating previous forecasts of a variable and the corresponding observations of the variable, in order to quantify the systematic forecast errors (bias) for each observation point. This bias results from many factors, including shortcomings in the physical-process parameterizations, and the inability of the model with a particular resolution to represent small-scale processes. The bias for each observation location is then used to correct future forecasts at the respective points. There are a number of MOS-type approaches, that differ in terms of the length of time over which the previous

Downloaded from https://www.cambridge.org/core. The University of British Columbia Library, on 14 Jan 2020 at 06:11:04, subject to the Cambridge Core terms of use , available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511763243.014

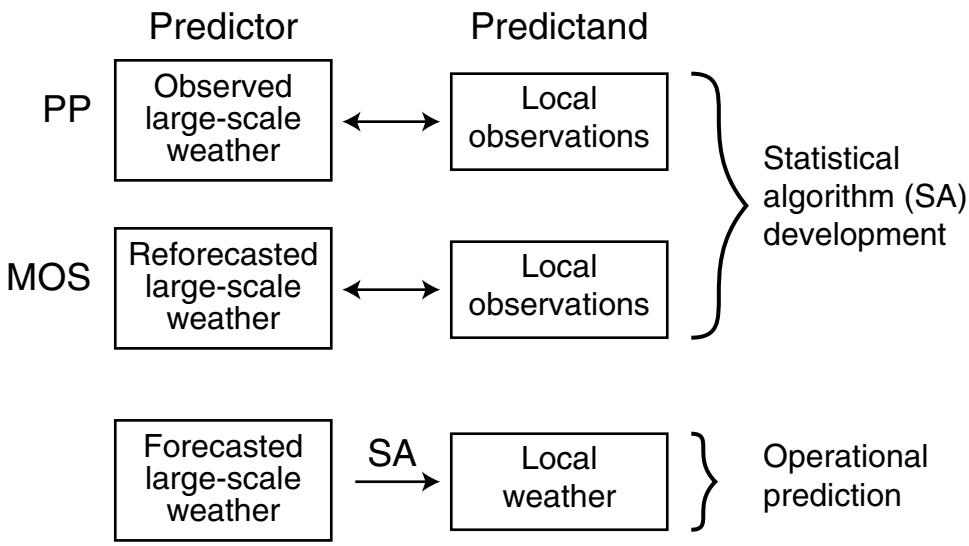

Fig. 13.2 Schematic of the PP and MOS approach to statistically relating variables or features that are reasonably well forecast by a model (predictors, left column) with those that are not (predictands, right column). With the PP method, historical archives of observations or analyses are used in the definition of the statistical relationship, whereas with the MOS approach, model reforecasts of historical cases are employed. The sections on MOS describe methods that use training periods of different length.

forecasts are used to define the bias. The classical, historical approach, referred to above as the static method, has been to calculate the statistical relationships over a multi-year historical period. The use of this long training period leads to stable statistics, but it is so computer-resource intensive that operational models cannot be frequently updated with improvements because that would invalidate the statistics (that would then need to be recalculated). As a result, other MOS-type methods with shorter training periods have been developed. An advantage of very-short training periods is that systematic errors can be weather-regime dependent, so adjustments based on recent model performance can be beneficial. Thus, a balance must be reached between (1) a short learning period that is vulnerable to missing data that are needed for training, and to the occurrence of extreme weather events with unrepresentative errors and (2) a long learning period that produces stable statistics but that is arguably too computationally expensive to be practical. The following sections review a few different MOS-based approaches to systematic-error reduction.

Conventional MOS

This approach, requiring statistics that are generated by forecasts from the same model over a period of at least two years, is summarized in Glahn and Lowry (1972). Because MOS requires the separate calculation of statistics based on forecast–observation pairs for each forecast lead time, for each observation location, and for each variable, a large number of equations are involved. Even though there has been a clear trend toward the use of shorter training periods with MOS-based methods, Hamill *et al.* (2004, 2006) present results that show that, for challenging situations such as long-lead-time forecasts, forecasts of rare events, or forecasts of surface variables with significant bias, long training periods

Downloaded from https://www.cambridge.org/core. The University of British Columbia Library, on 14 Jan 2020 at 06:11:04, subject to the Cambridge Core terms of use , available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511763243.014

can be beneficial. And Clark and Hay (2004) illustrate the great potential benefit of conventional MOS for producing improved forecasts for hydrological applications.

Jacks *et al.* (1990) provide a summary of an NCEP MOS system, where predictors included forecasts of temperature, temperature advection, thickness, precipitation amount, precipitable water, relative humidity, vertical velocity, horizontal wind components, wind speed, relative vorticity, vorticity advection, stability, and moisture convergence. These predictors are often defined at different levels in the model. The resulting system was very computationally demanding, involving the use of many thousands of statistical equations. It is interesting to note that in the late 1980s the MSC replaced its operational MOS system with PP products, which were used throughout the 1990s. See Brunet *et al.* (1988) for a discussion of the relative statistical characteristics of the PP and MOS methods.

Figure 13.3 shows an example of the benefit of the application of the conventional MOS approach, in this case in the context of mesoscale LAM simulations of 10-m AGL winds. The MM5 model was used for operational prediction during the 2002 Winter Olympics, for the Salt Lake City area, which is dominated by the complex orography of the surrounding area. The MOS equations were derived using three winter seasons of forecasts and observations for 18 mountain and valley locations. The grid increment of the model used for the generation of the statistics and for the operational prediction was 12 km, even though a 4-km nested grid was also employed in order to assess the benefit of higher horizontal resolution. The 4-km grid did not feed back to the 12-km grid, so it did not affect the MOS correction. However, the model version did change during the period of the MOS-equation development. The figure shows the wind speed MAE for both the 0000 UTC and 1200 UTC forecast cycles, based on the Direct Model Output (DMO) from the

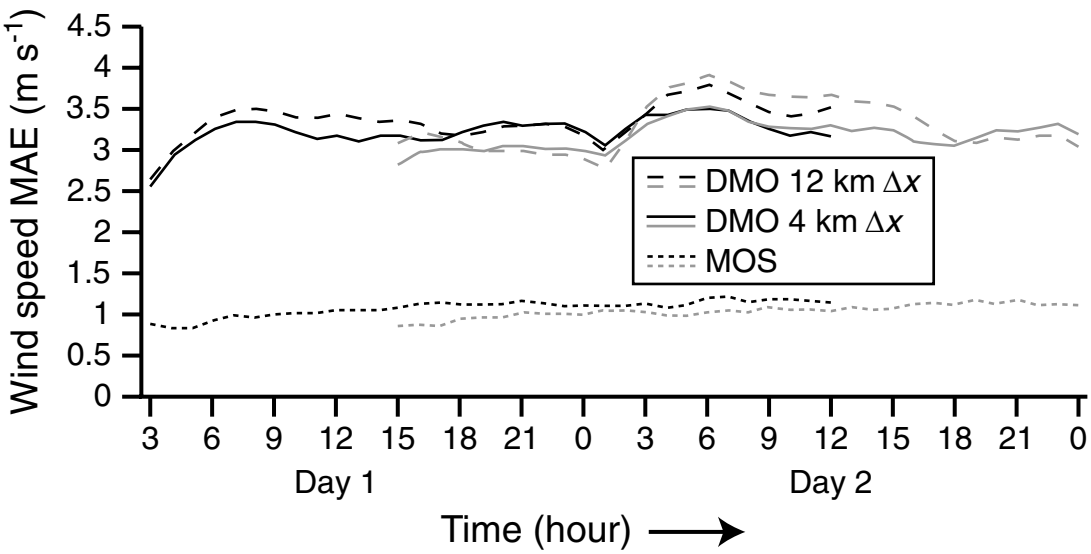

Fig. 13.3 The MAE of wind speed for 18 observation locations in complex terrain, based on DMO from MM5 LAM simulations that used horizontal grid increments of 12 km and 4 km, and MOS simulations that were based on the 12-km grid increment version of the model. The statistics for 36-hour forecasts from both the 1200 UTC (gray) and 0000 UTC (black) cycles are shown. Adapted from Hart *et al.* (2004).

Downloaded from https://www.cambridge.org/core. The University of British Columbia Library, on 14 Jan 2020 at 06:11:04, subject to the Cambridge Core terms of use , available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511763243.014

12-km and 4-km models, as well as based on the MOS from the 12-km model. Based on the four curves for the DMO, the higher horizontal resolution produced no significant benefit because much of the orographic variability was still subgrid-scale. However, the use of MOS reduced the average MAE from about 3.5 m s^{-1} to about 1 m s^{-1}.

Updatable MOS

Updatable MOS (UMOS, Wilson and Vallée 2002, 2003) allows frequent and automatic updating of statistical forecast equations soon after changes are made to the NWP model. This is accomplished through user-controlled weights, such that, after a model change is implemented, the statistical properties of the new model forecasts and those for the old model can be weighted and blended in the operational statistical relationship. That is, instead of training a new algorithm using a frozen version of the new model for a long period of time, independent of the operational system, with UMOS the statistical properties of the new model are gradually given more weight as a longer history is accumulated. In the MSC UMOS implementation of this system, the blending of the old and new systems begins after 30 cases from the new model have accumulated. After 300–350 cases with the new model have been included, the influence of the old model is neglected. Figure 13.4 compares the

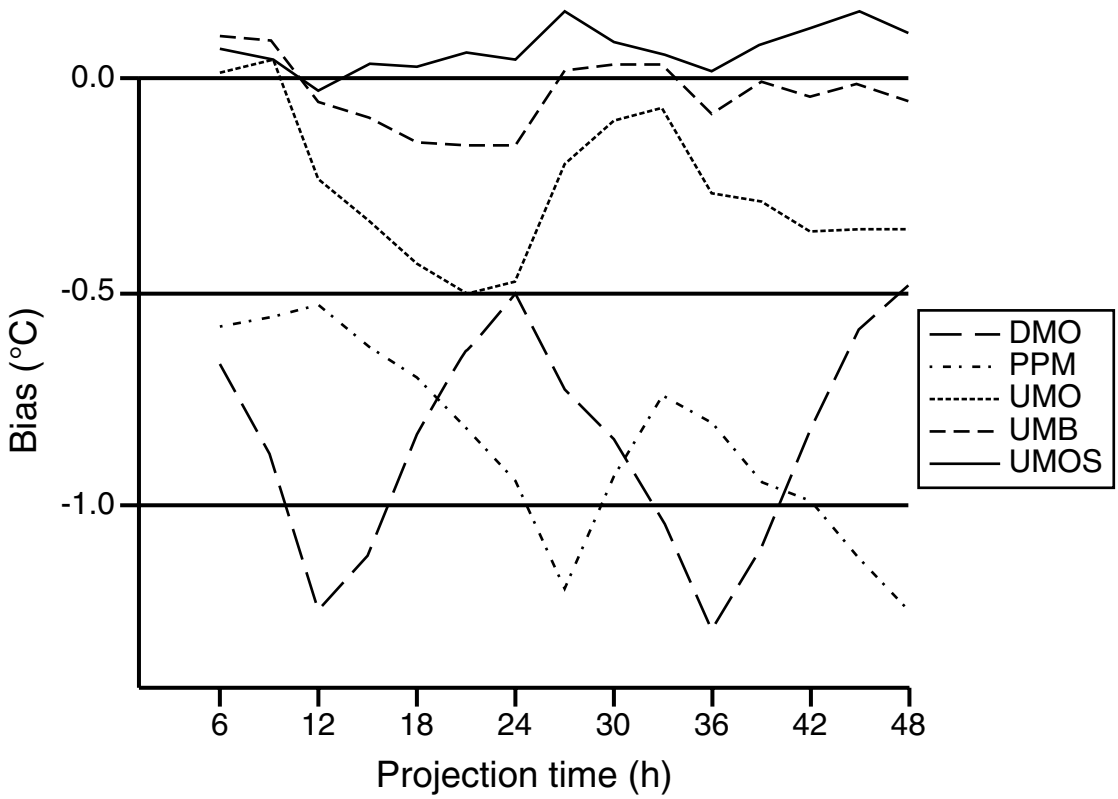

Fig. 13.4 Bias of winter-season temperature forecasts for about 250 Canadian stations, based on three statistical-correction methods related to Updatable MOS (UMOS, UMO, UMB) and the Perfect Prog Method (PPM). Also shown is the bias associated with uncorrected Direct Model Output (DMO). Adapted from Wilson and Vallée (2003).

Downloaded from https://www.cambridge.org/core. The University of British Columbia Library, on 14 Jan 2020 at 06:11:04, subject to the Cambridge Core terms of use , available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511763243.014

bias associated with a hierarchy of three applications of UMOS (UMO, UMB, complete UMOS) with that of the PP method and the statistically uncorrected DMO for 250 Canadian stations for a winter season. The UMO statistical equations contain no new-model data, so this method ignores the effect of the change in the model, using the old statistical equations with the new model. With UMB, some new-model data are used, but none that are close in time to the test period. The DMO has a negative bias that varies between –0.5 °C and –1.3 °C, depending on forecast lead time. Because the PP method does not correct for model biases, the PP method curve has a similar average bias. In spite of the fact that the UMO correction was based on old-model-version statistics, there is still significant bias correction. The complete UMOS approach has the smallest bias, averaged over all forecast lead times. See Wilson and Vallée (2002, 2003) for additional information about this method.

Very-short-update-period dynamic MOS

A review and evaluation of implementations of MOS with different training periods is provided in McCollor and Stull (2008c), where the model employed was the CMC GEM model (Côté *et al.* 1998a,b). Four bias-calculation approaches tested are summarized below.

- *Seasonal-mean error* – For cold-season forecasts, the average mean forecast error was calculated for the six-month period encompassing the previous cold season. Similarly, warm-season forecasts were corrected using errors from the previous warm season.
- *Moving average with uniform weighting* – The average mean forecast error was calculated using an unweighted average of the bias error from the previous *n* days.
- *Moving average with linear weighting* – Same as above, but using a linearly weighted average, with recent errors weighted more heavily.
- *Moving average with nonlinear weighting* – Same as above, but using a nonlinearly weighted average.

The objective of the weighting of course was to provide greater weight to the recent forecast errors, to be responsive to regime changes, while employing a significantly long averaging period to enhance statistical stability. Averaging windows from 1 to 24 days were evaluated in terms of their ability to reduce the forecast error. Figure 13.5 shows the MOS-adjusted errors in forecasted maximum temperature for the method that used the linear weighting. Each curve corresponds to a particular lead time within 8-day forecasts, and shows the error at that lead time as a function of the length of the different averaging periods involved in the calculation of the bias. For all lead times, the greatest incremental forecast improvement associated with adding days to the averaging period was for the shorter averaging times. The longer lead times benefited the most from the use of longer averaging windows. The 1- and 2-day lead time forecasts did not benefit much from the use of averaging windows of greater than 5 days, but the 8-day forecast benefited from the extension of the windows out to 15 or 20 days. Other studies have used error-weighting windows of 7 days (Stensrud and Skindlov 1996, Stensrud and Yussouf 2003), 12 days (Stensrud and Yussouf 2005), 14 days (Eckel and Mass 2005, Jones *et al.* 2007), 21 days (Mao *et al.* 1999), and 15–30 days (Woodcock and Engel 2005).

Downloaded from https://www.cambridge.org/core. The University of British Columbia Library, on 14 Jan 2020 at 06:11:04, subject to the Cambridge Core terms of use , available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511763243.014

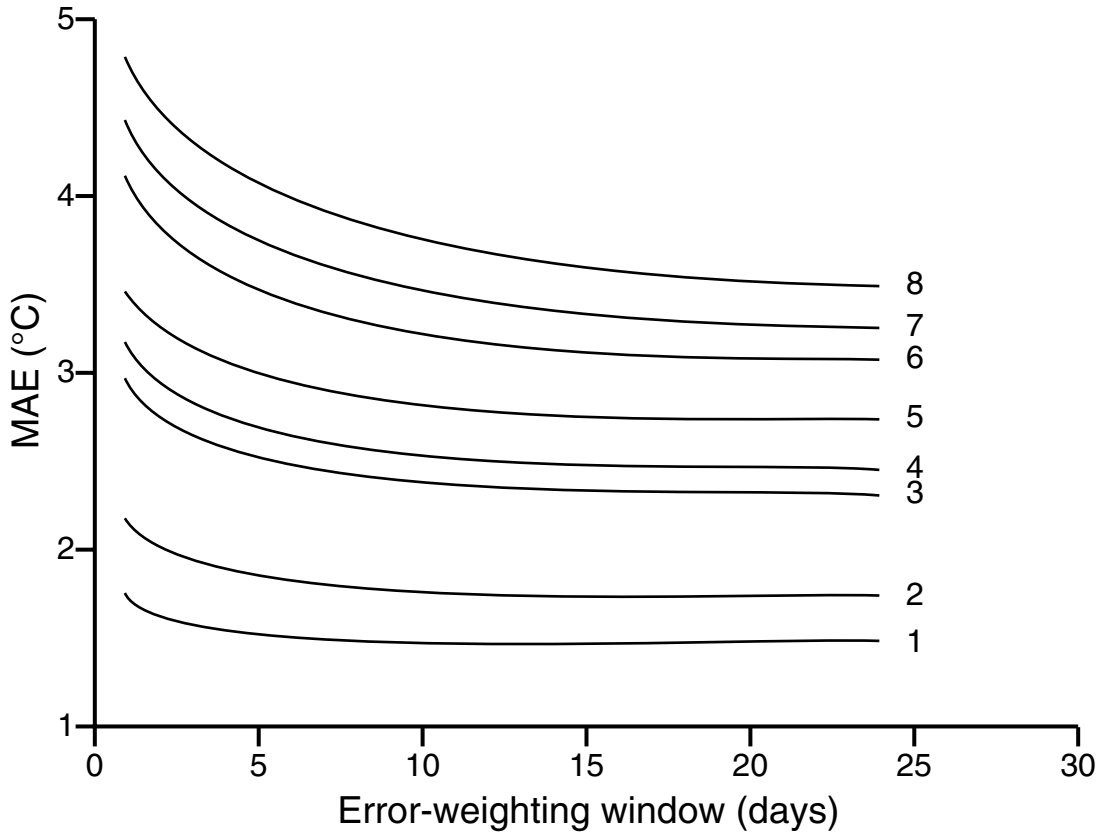

Fig. 13.5 Forecast errors (MAE) of maximum daily temperature (ordinate) for different MOS error-weighting windows (abscissa). Each curve corresponds to a forecast lead time (days). In the calculation of the bias correction used in a forecast, the biases from the previous *n* days (the window length) were averaged, using a linear weighting where the recent errors were weighted more heavily. Adapted from McCollor and Stull (2008c).

13.2.3 Kalman-filter methods

The use of Kalman filters (KF) is another automatic post-processing method for employing past observations and forecasts to estimate model bias in future forecasts. Delle Monache *et al.* (2006b) review the mathematical basis for the method. Analogous with Eq. 6.16, which describes the application of least-squares estimation with Kalman filtering to data assimilation, the following equation pertains to the bias-estimation problem:

$$B_{t+\Delta t} = B_t + \beta_t (y_t - B_t) .$$

The variable B is the estimate of the bias in some forecast variable. The quantity $B_{t+\Delta t}$ is the estimate of the bias in the variable at a forecast lead time Δt, B_t is the estimate of the bias at the end of the previous forecast, y_t is the observed forecast error (both

Downloaded from https://www.cambridge.org/core. The University of British Columbia Library, on 14 Jan 2020 at 06:11:04, subject to the Cambridge Core terms of use , available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511763243.014

systematic and random) at the end of the previous cycle (difference between the forecast and observations), and β is the Kalman gain. Say the forecasts are of 24 h duration (Δt), and we desire to estimate the bias in the forecast of B at this time ($B_{t+\Delta t}$) so that we can correct for it. For the forecast valid at the present time, the bias had previously been estimated to be B_t, and this is used as the first guess. Because the forecast valid at time t has completed, the total error y_t has been calculated. This is differenced with the previous estimate of the bias, and multiplied by the weighting factor β_t. Thus, the future bias is estimated to be the most-recent estimate of the bias that is adjusted by a weighted difference between this bias estimate and the observed total error. See Delle Monache *et al.* (2006b), and Appendix A therein, for a discussion of the calculation of the Kalman gain.

Delle Monache *et al.* (2008) illustrate the error reduction associated with the application of a KF procedure for each member of a multimodel ensemble of 24-h forecasts of ozone concentration (Fig. 13.6). Each of the first eight names in the legend corresponds to a particular photochemical model and meteorological model combination that was used in the construction of the ensemble. Position on the coordinate axes corresponds to the RMSEs associated with the systematic error (abscissa) and the random error (ordinate) of model forecasts of ozone concentration. The distance between the origin and any coordinate represents the total RMSE. The coordinate of the tail of each vector represents the RMSE values for the DMO, and the coordinate of the head of the vector represents the RMSE for the KF-corrected forecast. The direction and length of the vector show the amount of change in the systematic and random errors that results from the application of the KF correction. The "E" refers to the ensemble average of the forecasts, where the vector tail defines the RMSEs of the ensemble average of the DMO, and the head defines the RMSEs after the KF correction is applied to the ensemble average. Here, the ensemble averaging is done before the filtering. The tail of the "EK" vector applies to the ensemble average of the individual KF-corrected forecasts, and the position of the head results from the application of the KF correction a second time. Here, the filtering is done before the averaging. There was clearly a large decrease in the systematic error that resulted from the application of the KF to each of the ensemble members and to the ensemble mean. Note that this method does not require an extensive statistical database for training.

13.2.4 Gridded bias-correction

Statistical corrections based on standard MOS methods apply at observation points only, and thus it is not possible to straightforwardly infer the model bias in a more general way at any arbitrary location for which a forecast is desired. This ability to provide spatially distributed information about systematic error is important for many applications, such as when using forecast precipitation in a gridded hydrologic model, or when forecast temperatures interact with the land surface at every grid point to control evaporation and sensible-heat fluxes. Hacker and Rife (2007) show how computation of error covariance matrices can allow the definition of systematic error on a grid, and describe the implementation of the method in an operational LAM.

Downloaded from https://www.cambridge.org/core. The University of British Columbia Library, on 14 Jan 2020 at 06:11:04, subject to the Cambridge Core terms of use , available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511763243.014

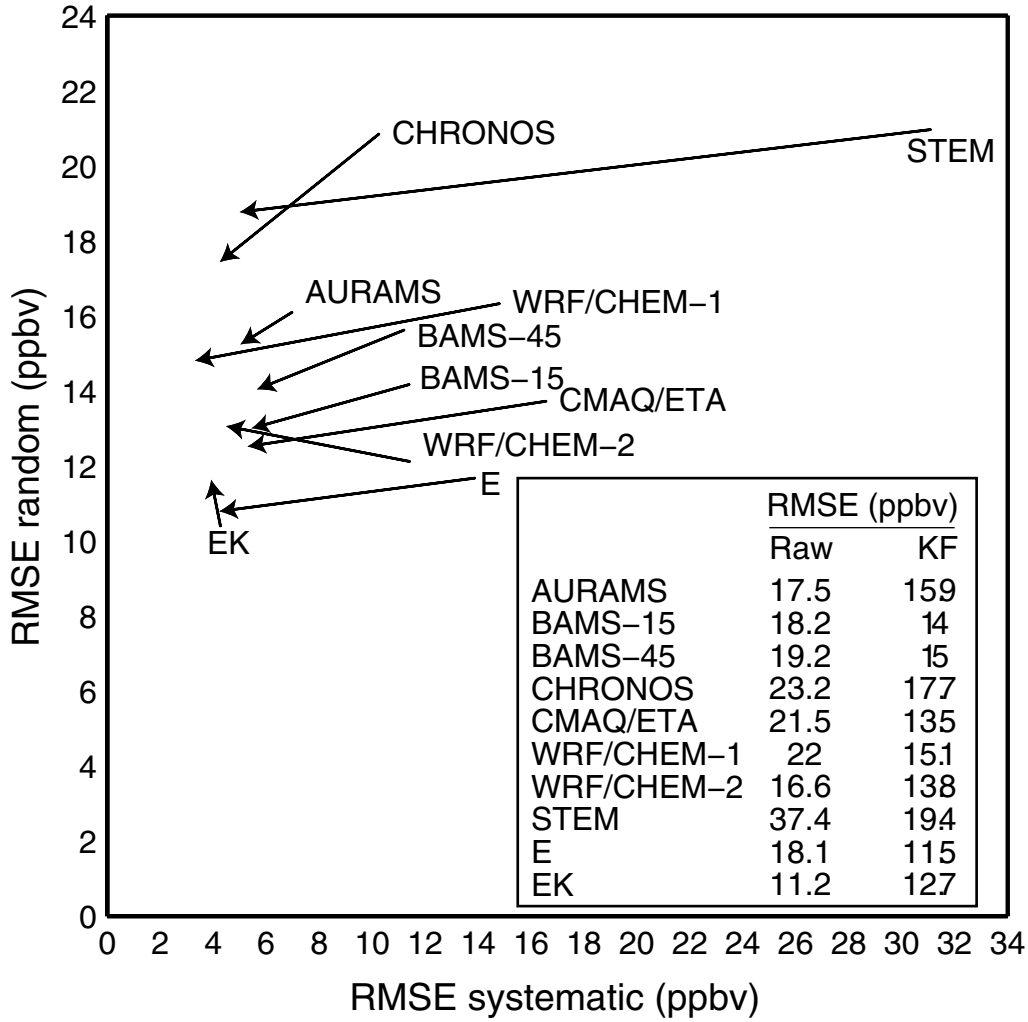

Fig. 13.6 The RMSE of DMO and KF-adjusted forecasts of ozone concentration for members of an ensemble. Each of the first eight names in the legend corresponds to a particular photochemical–meteorological model combination that was used in the construction of the ensemble, and "E" and "EK" refer to the application of the filtering and averaging in different orders (see the text). Position on the coordinate axes corresponds to the RMSEs associated with the systematic error and the random error of model forecasts of ozone concentration. The tail of each vector defines the RMSE values for the DMO, and the head represents the RMSE for the KF-corrected forecast. The direction and length of the vector show the amount of change in the systematic and random errors that results from the application of the KF correction. Adapted from Delle Monache *et al.* (2008).

13.3 Weather generators

Even though model time steps may be relatively short - perhaps tens of minutes – much of the short-time-scale variability associated with some phenomena is not represented in the model solution. For example, precipitation rates in nature can be highly variable, as rain

Downloaded from https://www.cambridge.org/core. The University of British Columbia Library, on 14 Jan 2020 at 06:11:04, subject to the Cambridge Core terms of use , available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511763243.014

bands, or other small convective features that are in various stages of their life cycle, pass across a location. Variability on these time scales is not represented in most operational models. This is especially true for large AOGCM grid boxes for which the time series of variables are smoother than those that apply to single points, simply because of the averaging that is implied over the large area. Unfortunately, high-frequency rain rates are needed for many hydrological applications, where the rate determines the partitioning of the rainwater between runoff and infiltration. Another example of high-frequency variability that is not represented in NWP models is wind gustiness, which is needed in models of dust elevation and transport, ocean waves, etc. To address such needs for high-frequency information from NWP and climate models, synthetic high-resolution time series can be generated with what are called *stochastic weather generators*. These methods essentially post-process the model-generated time series, adding realistic higher-frequency variability.

For NWP model simulations, the weather generators can add high-frequency spatial and temporal variations in the precipitation rate. For climate projections, which perhaps only have output at a monthly frequency, these generators can simulate the temporal distribution of wet and dry spells, the typical number of days with and without precipitation, etc. The generators can be tuned to apply to particular current and recent weather types. Information about the application of stochastic weather generators for climate-change studies, especially related to precipitation rate, can be found in Katz (1996), Semenov and Barrow (1997), Goddard *et al.* (2001), Huth *et al.* (2001), Palutikof *et al.* (2002), Busuioc and von Storch (2003), Katz *et al.* (2003), Wilby *et al.* (2003), Elshamy *et al.* (2006), Wilks (2006), and Kilsby *et al.* (2007).

Analogously, high-frequency wind-speed variability, sometimes known as gustiness or turbulence, is not represented in NWP or climate models, but it is important for predicting ocean-wave height, the elevation of dust from the surface in dust models, and dangers to aircraft. Application of weather generators for predicting gusts and turbulence will be discussed in Chapter 14, which deals with specialized models that are coupled to NWP models.

13.4 Downscaling methods

The concept of downscaling large-scale analyses and forecasts of weather and climate, such that small-scale features are estimated based on input about the larger-scale structure of the atmosphere, is discussed in Chapter 3 regarding the use of nested grids, and in Chapter 16 related to defining regional climates based on large-scale analyses or projections. The statistical downscaling of climate simulations, from interseasonal to century time scales, is described in Section 16.3.1, and has much in common with the MOS-based statistical methods described above.

SUGGESTED GENERAL REFERENCES FOR FURTHER READING

Hamill, T. M., J. S. Whitaker, and S. L. Mullen (2006). Reforecasts: An important data set for improving weather predictions. *Bull. Amer. Meteor. Soc.*, **87**, 33–46.

Downloaded from https://www.cambridge.org/core. The University of British Columbia Library, on 14 Jan 2020 at 06:11:04, subject to the Cambridge Core terms of use , available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511763243.014

McCollor, D., and R. Stull (2008c). Hydrometeorological accuracy enhancement via post-processing of numerical weather forecasts in complex terrain. *Wea. Forecasting*, **23**, 131–144.

Wilks, D. S. (2006). *Statistical Methods in the Atmospheric Sciences*. San Diego, USA: Academic Press.

PROBLEMS AND EXERCISES

1. Speculate on the possible sources of systematic and random errors in NWP-model forecasts, in addition to those listed in this chapter. Distinguish between the two sources, and if necessary explain why the error is in one category or the other.
2. In reference to Fig. 13.1, why might the wind have a larger percentage of the error associated with the random component than does the temperature?
3. Again in reference to Fig. 13.1, describe situations that could be responsible for the considerably larger error at some locations compared to others.
4. Why might there be a greater need for statistical correction of model error for levels near the ground?

Downloaded from https://www.cambridge.org/core. The University of British Columbia Library, on 14 Jan 2020 at 06:11:04, subject to the Cambridge Core terms of use , available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511763243.014