An Evaluation of Analog-Based Postprocessing Methods across Several Variables and Forecast Models

BADRINATH NAGARAJAN, LUCA DELLE MONACHE, AND JOSHUA P. HACKER
National Center for Atmospheric Research, Boulder, Colorado

DARAN L. RIFE
GL Garrad Hassan, San Diego, California

KEITH SEARIGHT
Cooperative Institute for Research in the Atmosphere, Colorado State University, Fort Collins, Colorado

JASON C. KNIEVEL
National Center for Atmospheric Research, Boulder, Colorado

THOMAS N. NIPEN
University of British Columbia, Vancouver, British Columbia, Canada

(Manuscript received 12 July 2014, in final form 11 August 2015)

ABSTRACT

Recently, two analog-based postprocessing methods were demonstrated to reduce the systematic and random errors from Weather Research and Forecasting (WRF) Model predictions of 10-m wind speed over the central United States. To test robustness and generality, and to gain a deeper understanding of postprocessing forecasts with analogs, this paper expands upon that work by applying both analog methods to surface stations evenly distributed across the conterminous United States over a 1-yr period. The Global Forecast System (GFS), North American Mesoscale Forecast System (NAM), and Rapid Update Cycle (RUC) forecasts for screen-height wind, temperature, and humidity are postprocessed with the two analog-based methods and with two time series–based methods—a running mean bias correction and an algorithm inspired by the Kalman filter. Forecasts are evaluated according to a range of metrics, including random and systematic error components; correlation; and by conditioning the error distributions on lead time, location, error magnitude, and day-to-day error variability.

Results show that the analog methods are generally more effective than time series–based methods at reducing the random error component, leading to an overall reduction in root-mean-square error. Details among the methods differ and are elucidated upon in this study. The relative levels of random and systematic error in the raw forecasts determine, to a large extent, the effectiveness of each postprocessing method in reducing forecast errors. When the errors are dominated by random errors (e.g., where thunderstorms are common), the analog-based methods far outperform the time series-based methods. When the errors are strictly systematic (i.e., a bias), the analog methods lose their advantage over the time series methods. It is shown that slowly evolving systematic errors rarely dominate, so reducing the random error component is most effective at reducing the error magnitude. The results are shown to be valid for all seasons. The analog methods show similar performance to the operational model output statistics (MOS) while showing greater reduction of random errors at certain lead times.

1. Introduction

Forecasts from numerical weather prediction (NWP) models are subject to errors from several sources, including initial conditions, lateral boundary conditions, and the models’ formulations (e.g., numerical truncation and parameterization of unresolved physical processes).
Errors in initial conditions are sometimes addressed through the use of an ensemble of initial conditions (e.g., Magnusson et al. 2008; Bowler 2006; Kalnay 2003). For limited-area models, errors in lateral boundary conditions are handled by using perturbed boundary conditions (Saito et al. 2012). On the other hand, model errors are addressed through the use of accurate numerical schemes, a stochastic kinetic energy backscatter scheme (Berner et al. 2011), and stochastic physics (Buizza et al. 1999). All of these approaches improve NWP model forecasts. However, forecasting near-surface variables (e.g., 10-m wind speed) remains a challenge as a result of a poor representation of atmospheric boundary layer processes (e.g., stable stratification) in NWP models.

Additional increases in the accuracy of NWP forecasts can be achieved through postprocessing methods (e.g., Glahn and Lowry 1972; Kalman 1960; Crochet 2004; Leith 1978; Stensrud and Skindlov 1996). An objective of a postprocessing method is to improve the accuracy of the forecast by reducing both systematic and random errors, while preserving or improving the correlation with observations (Wilks 2006).

Analog-based methods are one means of post-processing. They have been applied, for instance, to long-range weather prediction (Bergen and Harnack 1982), short-term visibility and mesoscale transport forecasts (Esterle 1992; Carter and Keisler 2000), and medium-range precipitation predictions (Hamill et al. 2006). Recently, Delle Monache et al. (2011) introduced two new analog-based methods. Both chose analogs by calculating a Euclidean distance between a current forecast and a history of forecasts at an individual observing site. From there, the first method (herein abbreviated AN) forms an ensemble from the observations that correspond to the best analogs. A deterministic forecast is simply the weighted mean of the ensemble, but the distribution can also be used to produce probabilistic forecasts (Delle Monache et al. 2013). In a second, variant method (herein abbreviated ANKF), a Kalman filter predictor–corrector algorithm (e.g., Bozic 1994; Roeger et al. 2001; Delle Monache et al. 2006, 2008) is applied to the analogs arranged into a series that is rank ordered by descending distance to the current forecast. The result is a corrected raw deterministic forecast whose error estimate is based on an exponential weighting of past errors with the closest analogs given the most weight, contrasting with the linear average in the AN.

For 400 surface stations in the western United States, Delle Monache et al. (2011) applied AN and ANKF to 6 months of 10-m wind speed forecasts from the Weather Research and Forecasting (WRF) Model. The AN and ANKF methods significantly reduced the bias and random errors in the WRF forecasts, yielding an accurate prediction. AN and ANKF also improved the correlation between the corrected forecast and observations as characterized by the rank correlation. Furthermore, the two methods yielded improved forecasts of the wind speed under rapidly evolving weather regimes, a forecasting challenge for NWP models.

In this study, we investigate the robustness and generality of the analog-based methods’ performances by extending them to (i) the entire conterminous United States (CONUS); (ii) forecasts from the North American Mesoscale Forecast System (NAM; Janjic et al. 2010), Global Forecast System (GFS; information online at http://www.emc.ncep.noaa.gov/GFS/doc.php), and Rapid Update Cycle (RUC; Benjamin et al. 2004); (iii) the near-surface variables 10-m wind speed, 2-m temperature $T$, and 2-m relative humidity $RH$; and (iv) a full year of verification. The study addresses influences from the error properties of the NWP model to be postprocessed, the region in the United States, and the forecast lead time. Results help explain when and where analog methods are expected to be successful. Comparison against a 7-day running mean (herein abbreviated 7-Day) and a Kalman filter predictor corrector (herein abbreviated KF), which are both time series error filtering–based methods, also provides new context for understanding how the analog methods perform. The performance of the different postprocessing methods for different variables lends insight into some details of error characteristics that differ among those variables. Finally, for the benefit of the operational forecasting community, the postprocessing methods are compared against model output statistics (MOS) from GFS and NAM.

Section 2 describes the data, and briefly summarizes AN and ANKF. Section 3 contains the results from AN and ANKF compared to 7-Day and KF and presents the results as a function of the season. The postprocessing methods are also compared against the operational MOS. Finally, section 4 provides a summary and conclusions.

2. Data and methods

We used GFS, NAM, and RUC predictions from the 12-month period starting on 1 August 2010. The last two months, June and July 2011, are reserved for the verification presented herein. To exploit all of the available data and mimic real-world forecast operations, the training period was increased from 10 months for the first forecast verified, which was initialized on 1 June 2011. Training data accumulate through the 2-month verification period, so that 12 months are available for
the last forecasts (initialized on 31 July 2011). The seasonal dependence of the methods’ performance is analyzed by extending the verification to an additional five 2-month periods covering the rest of the year.

a. Data

Figure 1 shows the spatial distribution of 740 surface stations over the CONUS. The stations were chosen based on the availability of high-quality observations of \( T \), RH, and wind speed over the CONUS. Stations reporting less than 85% of the time over the 1-yr period were rejected. Choosing a threshold greater or less than 85% gives relatively less or more dense coverage east of 100\(^{\circ}\)W, respectively. The observations came from a quality-controlled Meteorological Assimilation Data Ingest System (MADIS; information online at http://madis.noaa.gov). Miller and Benjamin (1992) provided a detailed description of the quality control procedure used in MADIS. The procedure checks for spatial (i.e., buddy checks) and temporal continuity (i.e., rate of change of a variable with time), and ensures that the variable falls within a reasonable range of values.

The postprocessing methods are applied to NAM, GFS, and RUC forecasts initialized daily at 1200 UTC. NAM and RUC (Janjic et al. 2010; Benjamin et al. 2004) are operational mesoscale models providing coverage over the CONUS, while GFS is a global model. Each provides predictions of variables used as input into the postprocessing methods. Table 1 lists the input data and characteristics. The analog method works best when the forecast dataset used for training the algorithm is produced by a frozen version of the model. Operational models undergo upgrades routinely. This results in forecast errors that are less representative of the current forecasts. The GFS and RUC models did not undergo significant upgrades between August 2010 and July 2011. However, NAM was upgraded in March 2011, which significantly impacted the temperature forecasts with no significant impacts on the wind speed and dewpoint forecasts. Thus, the temperature results for NAM must be interpreted by taking into account the model upgrade.

The MOS data from GFS and NAM for the months of June and July 2011 (data online at http://www.mdl.nws.noaa.gov/~mos/archives/met.html for NAM and http://www.mdl.nws.noaa.gov/~mos/archives/mav.html for GFS) are compared against the postprocessing methods.

Delle Monache et al. (2013) successfully applied AN to 0000 UTC initialized daily forecast data. Thus, results

![Figure 1. The blue circles depict the spatial distribution of 740 surface stations over the CONUS.](image)

The MOS data from GFS and NAM for the months of June and July 2011 (data online at http://www.mdl.nws.noaa.gov/~mos/archives/met.html for NAM and http://www.mdl.nws.noaa.gov/~mos/archives/mav.html for GFS) are compared against the postprocessing methods.

Delle Monache et al. (2013) successfully applied AN to 0000 UTC initialized daily forecast data. Thus, results

| TABLE 1. NWP forecast data from GFS, NAM, and RUC. Variables \( P_{\text{surface}} \), PBLH, \( T \), UWLW, and RH represent the surface pressure, planetary boundary layer height, 2-m \( T \), surface upwelling longwave radiation, and 2-m RH, respectively. The predictor variables from the forecast model output data are interpolated to station locations and are explained in section 2b. |
|----------------|---------|---------|
|                | GFS     | NAM     | RUC     |
| Horizontal resolution | 0.5° × 0.5° | 12 km   | 13 km   |
| Forecast length used (h) | 72      | 72      | 12      |
| Forecast output frequency (h) | 3       | 3       | 1       |
| Predictor variables for 10-m \( V \) (all models) | Speed, direction, \( P_{\text{surface}} \), PBLH, and \( T \), direction, and UWLW |
| Predictor variables for 2-m \( T \) (NAM) | \( T \), direction, and UWLW |
| Predictor variables for 2-m RH (NAM) | RH, direction, \( T \), and PBLH |
reported here should be applicable to 0000 UTC forecasts from NAM, GFS, and RUC.

b. Methods

Postprocessing methods relate the predictors to a particular forecast parameter of interest (a predictand). Each postprocessing method here has different properties and brings with it different expected advantages and disadvantages. The time series–based methods, namely 7-Day and KF, are used for a reference to compare the analog-based methods, following Delle Monache et al. (2011). The 7-Day is chosen for its simplicity. The KF is chosen for its ability to adapt quickly to rapid changes in forecast errors associated with rapidly evolving weather systems, and for its widespread use as a postprocessing method. A brief description is provided of each postprocessing method.

In 7-Day (Stensrud and Skindlov 1996), the mean of the error over the past 7 days, at a particular local time of day, gives an error prediction for a current forecast. The 7-day period is approximately the synoptic time scale, and the error predictions are effective on that scale. A key drawback to 7-Day is that the filter length is such that the error predictions cannot change at shorter time scales (i.e., when errors change rapidly).

KF is a predictor–corrector method that estimates the current error as an exponentially weighted average, with the most recent errors being given more weight than the older errors (see the appendix). Using the error estimate, a predicted correction is applied to a current forecast. KF can be trained quickly and adapts to changing errors. The exponential weighting results in somewhat less susceptibility to failure at error time scales between 2 and 7 days, but it is still unable to change error predictions on a daily time scale (or more correctly, at the fastest scale in a set of discrete data). In particular, if the current forecast has errors different from the errors over the most recent time history, the error prediction can be poor, and the postprocessed forecast might not be more accurate than the raw forecast.

Analog methods are designed to condition the training data based on similarity to the current forecast, as opposed to proximity in time. In AN, the forecast is determined from a weighted average of observations that occurred at the times when the best analog forecasts were valid. Several sensitivity tests performed by Delle Monache et al. (2011) suggest the use of 10 observations that verified at the time of the 10 best analogs. The weights assigned to the observations are inversely proportional to the square of the metric described for the AN method [Eq. (6); Delle Monache et al. (2011)]. Because direct use of analogs is not a filtering process, the ability to adapt to rapid changes in the weather is not restricted.

In the variant, ANKF (Delle Monache et al. 2011), analogs are arranged in descending order in the analog distance metric (i.e., from farthest to closest). In practice, the analog search algorithm rank orders the entire training set. All data are used in the ANKF, but the weighting is such that poor analogs receive negligible weight. KF is then run through the ordered set of analogs to estimate the current forecast error, which is an exponential weighting of the errors, with the closest analogs given the most weight. See Delle Monache et al. (2011) for further details. Because ANKF estimates and predicts the error, rather than directly using the observations as in AN, it can be superior to AN when the best analogs have errors most like the current forecast. Conversely, ANKF is more susceptible to failure when the best analog, or analogs, have different errors than the current forecast. The latter can occur when any component of the total forecast error is random, or when ignorance of additional predictors limits the skill in selecting analogs.

The distance metric for choosing analogs is directly proportional to the difference between a current forecast and a past prediction, and with the past forecasts exhibiting a similar temporal trend as the current forecast [for details see Eqs. (7) and (8); Delle Monache et al. (2011)]. Following Delle Monache et al. (2011), the predictors are chosen through trial and error.

c. Metrics

Verification metrics are chosen to reveal the levels of systematic and randomlike error in the forecasts, and to what extent postprocessing mitigates those errors. We apply four error metrics to the postprocessed forecasts and to the raw forecasts from the NWP models. Each offers a different, complementary view on the errors. Conditioning the verification data based on forecast time or location allows for interpretation.

Basic metrics include the mean absolute error (MAE), bias, centered root-mean-square error (CRMSE), and rank correlation (RC). For an individual forecast time and location, the error is \( \epsilon = f - o \), wherein \( f \) is a forecast and \( o \) is the verifying observation. Then, \( \text{MAE} = |f| \) and the overbar denotes an average over a sample. Bias and CRMSE arise from decomposition of RMSE. Here, we use the term bias to refer to a mean error over a given dataset. That is, \( \text{bias} = \bar{\tau} \). It is simply the mean error over the sample and is systematic. The remaining error is considered random, measured by \( \text{CRMSE} = \sqrt{(\bar{\tau}^2 - \epsilon^2)} \), also called the standard error. Those two metrics can be evaluated independently, or can be combined to give a geometric interpretation that arises from the well-known
decomposition of the RMSE (e.g., Taylor 2001). Finally, monotonic association between two variables can be addressed with the Spearman RC. It is more general than the linear Pearson correlation because it allows for nonlinear relationships between predictions and observations.

3. Results

This section presents quantitative comparisons and offers interpretations that guide when and where each postprocessing method might prove most helpful in correcting for forecast errors. The forecast models NAM, RUC, and GFS offer different forecast lead times and resolve different scales of motion, resulting in different error characteristics.

Improvements realized from postprocessing might depend on the physical location and time of day, for example. The underlying cause for varying improvements is variability in the magnitudes of random and systematic error. To aid interpretation, Fig. 2 shows the CRMSE and bias for some of the NWP forecasts, prior to postprocessing. In all cases, the random error component, as measured by CRMSE, is greater in magnitude than the systematic component, as measured by bias. Wind speed biases in the NAM are small, while the GFS forecasts exhibit a diurnal variability (Fig. 2a), with a slow bias late in the afternoon and a fast bias early in the morning prior to sunrise. The RUC shows a jump in the wind speed bias during the first 3 h, which may be related to an adjustment from background fields, and an improvement from a slow bias to near zero through the 12-h forecast that ends near sunset (Fig. 2b). The NAM temperature and RH forecasts show slowly increasing biases with a superimposed diurnal variability (Figs. 2c,d), of which the maxima are near midnight and the minima near midday. NAM temperature and RH CRMSE vary in opposite phase to the biases. CRMSE is low during the day and higher at night. NAM shows less random wind speed error (CRMSE) than does the GFS, with the highest value late in the day and the lowest value early in the morning prior to sunrise.

a. Dependence on forecast-error magnitude and rapidly changing errors

Here, we verify whether the expected behavior of the four postprocessing methods holds over a wide range of spatial locations and times of day. The magnitude of a current forecast error, or the difference between a current forecast error and the most recent forecast error in...
the training set, can help elucidate some of the different behaviors.

The simple 7-Day will work well only when errors evolve on time scales as long or longer than the synoptic time scales. AN is expected to minimize the influence from an error in the current forecast that is a large departure from errors in the most recent forecasts, because the observations corresponding to the analog only depend on forecast similarity in the entire training set. KF is expected to perform well when the current forecast error is similar to the most recent forecasts. ANKF is expected to perform well when the closest analogs have errors similar to the current forecast error. Here, we compare each method to the error in the raw forecast, binned by error magnitude and daily error change.

1) WIND SPEED

Binning and averaging the errors according to the MAE for a given forecast time (i.e., the average is over all available observations) shows dependence on error magnitude. Here, the MAE is computed for each observing location and lead time, with a sample formed from all forecasts (2 months). Those MAE distributions are binned, and the associated improvement from postprocessing is computed for the observations and lead times in each bin.

Results show that all postprocessing methods improve the GFS wind speed forecast at all error magnitudes (Fig. 3). All postprocessing methods improve the NAM forecast, except for 7-Day at large NAM MAE. Presumably, these cases of large errors in the NAM are not preceded by similar large errors during the previous 7 days. KF is an improvement compared to 7-Day for both GFS and NAM. This results from the ability to adapt more quickly to errors on shorter time scales. Both AN and ANKF give more accurate forecasts than KF, showing that the analogs provide corresponding observations that are better predictors than the recent error history provides.

Although the improvement grows with error magnitude almost monotonically, the improvement relative to the error magnitude is at best constant for GFS and decreases for NAM. This is consistent with the finer scales in NAM adding to the random error component, as a proportion of total error, compared to the coarser GFS.

Further insight comes from binning by the daily change in forecast error. Daily change in forecast error is defined as $|\text{MAE}_d - \text{MAE}_{d-1}|$, wherein $d$ is day, computed from the raw NWP forecast. The percent improvement in MAE for day $d$, from postprocessing, as it varies by the daily change in NWP prediction error, is shown in Fig. 4.

The relative improvement from KF and 7-Day with respect to GFS starts at about 40% for small day-to-day changes in the forecast-error magnitude (Fig. 4a). However, as $|\text{MAE}_d - \text{MAE}_{d-1}|$ grows, the improvements diminish, and 7-Day diminishes the forecast accuracy when the daily change in error is large. This expected (Delle Monache et al. 2011) result is somewhat mitigated by the KF, which offers no benefit for large error changes, but avoids making the forecasts worse. This is possible when the most recent days have errors closer to those in the current forecast, compared to the errors in the forecasts from 7 days prior.

Figure 4b shows the improvement over NAM forecasts by each postprocessing method. The maximum improvement by any method is about 30% (compared to greater than 45% for GFS) for small daily variability; the 7-Day and KF degrade the NAM forecast when the daily error changes are greater than 2.25 and 3.25 m s$^{-1}$, respectively. These results are again consistent with greater randomness in NAM errors.
The upward-trending curve of AN on NAM also shows the effects of random errors in the NAM. For small variation of $j_{MAE_d}$ between 0.25 and 1.25 m s$^{-1}$, AN offers less improvement to NAM than do the other methods. But when the error change is greater than 2 m s$^{-1}$, AN performs best. This result, similar to what Delle Monache et al. (2011) found, shows most clearly the distinction between time series–based methods and analog-based methods. Time series–based methods such as 7-Day and KF are effective at handling persistent errors. Analogs distributed evenly around a current forecast will, for small errors, give a relatively even distribution of historical observations, and the resulting error estimation can be slightly smaller in magnitude than from the time series–based methods.

The ANKF’s benefit drops as day-to-day error variability increases, but remains positive. The fact that it predicts the error and does not directly use observations causes the drop. The observations corresponding to the best analogs can be distributed around the verifying observation for day $d$, while the best analogs can have a systematic error that is different from the error in the current forecast. The ANKF results (Figs. 4a,b) show that this occurs for both the GFS and NAM forecasts. For NAM, the ANKF outperforms AN for small daily variability in errors. Small error variability in the NAM means that errors in the good analogs can predict the error in the current forecast better than the historical observations can predict the verifying observation. Results from RUC are qualitatively similar to results from NAM, owing to the models’ similar error characteristics.

The bin counts in Figs. 3 and 4 show that large errors and large daily changes in error are more rare than are small errors and small daily changes in error. This is of course expected. But it shows clearly that AN and ANKF can provide benefit for more rare error events.

2) SURFACE AIR TEMPERATURE AND RELATIVE HUMIDITY

Temperature and RH can have different error characteristics. Comparing these to the error characteristics in wind speed from NAM demonstrates more attributes of the various postprocessing methods. First, improvements to NAM 2-m surface air temperature and RH increase with MAE (Fig. 5). Different from the wind speed, the level of improvement drops at the largest levels of MAE, although the large variations indicate sampling error, so that result might not be meaningful. Here, KF performs as well as AN and 7-Day does nearly as well. This indicates that systematic temperature errors have a component that exhibits a time scale slower than the synoptic time—that is, more like a bias, which can be exploited by any method here. AN loses its advantage when the best analogs do not necessarily correspond to observations that are distributed around a verifying observation on a day with very large errors in the NAM.

Different than for wind speed, AN for $T$ does not show a clear advantage when forecast errors are large. This result can occur when the error is large and the forecast is also in the tails of the climatology of a model. Consider the case of a constant error (i.e., the limit of an infinite error time scale) in a model and a rare forecast. The best analogs will not be distributed around the current forecast in any way resembling a normal distribution. Instead, they will be highly skewed toward the mean of the model climate. Assuming the model is skillful (although biased), then the corresponding observations going into AN will also be highly skewed toward the mean and will not represent the verifying observation for the current forecast. Instead, KF and
ANKF, which estimate the constant error, will perform perfectly. For $T$ in the NAM, it appears that the systematic nature of the error is sufficiently large that when a rare weather event occurs with a large forecast error, AN can be at a disadvantage. The frequency of this could be measured a priori to determine when AN will fail. The bin counts in Fig. 5 suggest that failure is very uncommon. For wind speed (Fig. 3b), relatively larger errors can occur when the forecast is not as rare an event, leading to a clearer advantage for AN.

The more randomlike nature of RH error favors AN and ANKF over the other methods (Fig. 5b). All the methods deteriorate somewhat at higher error levels, similar to NAM $T$ forecasts, but all maintain some benefit (ignoring the largest and most rare errors). Following the thought experiment above, systematic error at high daily error levels is less relevant and harder to predict, and the analog distributions are skewed somewhat. Still, accuracy improvements are possible.

Temperature and RH forecast errors, binned by the daily error variability, are consistent with errors in wind speed. Figures 6a and 6b show the improvement over NAM $T$ and RH. ANKF offers the most improvement over NAM for smaller day-to-day variations in forecast error; AN offers the most improvement for larger variations. Again, the observations associated with the analogs predict the verification for the current forecast, and the daily change in the error challenges the error-prediction methods.

This section demonstrated differences among methods depending on how the forecasts and errors behave when the errors are averaged in space and aggregated across all forecast lead times. To improve our understanding of the trade-offs between random and systematic errors, we turn to geometric interpretations of the bias and CRMSE.

b. Bias and CRMSE

For our purposes, bias is the mean error over all stations and at all lead times. CRMSE is the corresponding standard deviation of the error. Figures 7a and 7b show bias against CRMSE for wind speed. In Figs. 7a and 7b, the distance of a point on the curve from the origin is equal to RMSE, which is reported numerically next to each marker. Each marker represents an NWP forecast or a postprocessed forecast (i.e., 7-Day, KF, AN, and ANKF). All the methods reduce the bias and CRMSE of GFS (Fig. 7a) and RUC (Fig. 7b). AN and ANKF reduce CRMSE the most, followed by KF and 7-Day. Interpreting CRMSE as the random error component, the result is consistent with the last section. That is, over the broad range of NWP errors and their daily variability, there is benefit to conditioning the training data based on analog selection rather than time. The result holds for the NAM forecasts as well (not shown).

Figures 8a and 8b show the variation of bias with CRMSE for $T$ and RH resulting from different post-processing applied to the NAM forecasts. All the methods reduce the RMSE in the forecasts. The trade-off between random and systematic error components is clear. Unlike wind speed and RH, KF and 7-Day appear to reduce the bias in $T$ (Fig. 8a) to closer to zero than AN does. The time series methods clearly reduce the bias. Although AN reduces CRMSE slightly more than KF does, the bias reduction makes it as effective at reducing overall RMSE as the analog methods. For RH as well (Fig. 8b), the time series methods reduce the bias more, while the analog methods are more effective at reducing the CRMSE.

So far the samples for verification have been binned by error, change in error, or not at all. Because errors arising from chaotic error growth, which appear
random, can depend on forecast lead time, we might expect the benefit from postprocessing to vary in time. We next evaluate whether this is the case.

c. Dependence on forecast lead time

To explore how the behavior of each method depends on forecast lead time, we estimate confidence intervals using 1000 bootstrapped samples on the distributions of error differences, with replacement (Efron and Tibshirani 1994). Gilleland (2010) highlighted that when comparing error metrics from two methods, an overlap of confidence intervals might suggest no significant difference. To correctly infer the performance difference between two methods, the confidence interval should be computed on the difference of the error metric (e.g., bias) between the two methods. Thus, in this section, the performance is evaluated using the difference in CRMSE, RC, and the bias between an NWP forecast and the various postprocessed forecasts. Confidence intervals indicate whether the differences are expected to be different from zero.

1) WIND SPEED

Both the CRMSE and the rank correlations of NAM, GFS, and RUC can be improved significantly with postprocessing (Fig. 9). In these plots, a positive difference indicates an improvement over the raw forecasts. A diurnal variation is evident in each, except RUC, which...
is only integrated for 12 h and cannot show it. The shaded areas in Fig. 9 show the 95% confidence intervals derived from bootstrapping. Figures 9a–c exhibit a significant reduction in the CRMSE by all the methods except 7-Day applied to RUC (Fig. 9c). In the case of NAM and GFS, all the methods exhibit maxima in CRMSE difference (Fig. 9a) with the bias difference (Fig. 10a) during the night shows that the maximum reduction of

Minima in the improvement from postprocessing are around 6, 30, and 54 h (midday) for NAM (Fig. 9a) and are similar for GFS (Fig. 9b). In the case of NAM and GFS, all the methods exhibit maxima in CRMSE reduction between forecast hours 12–24, 36–48, and 60–72 h (Figs. 9a,b), during the night when NAM CRMSE is at a relative minimum (Fig. 2a). We cannot judge the daily behavior of RUC because of the short forecast time. NAM and GFS (Figs. 9a,b) exhibit similar behavior, which may arise from the NAM model deriving its lateral boundary conditions from the GFS forecasts.

The temporal evolution of RC improvement is shown in Figs. 9d–f. Again, the AN method performs best, mainly because RC also measures the random component of the error. For NAM and GFS, all methods again show a diurnal cycle (Figs. 9d,e) with maxima occurring near local sunset (forecast hours 12–15, 36–39, and 60–63). These times correspond to relatively high RC in the NAM forecasts, while GFS exhibits low RC throughout (not shown, and likely a result of the coarse resolution). The improvement in RC for GFS (Fig. 9e) is much larger than for NAM (Fig. 9d) because of the poor correlation in the raw forecasts. Seidel et al. (2012) compared planetary boundary layer (PBL) height $Z_h$ from ERA-Interim (ERA-I) and two PBL parameterization schemes against observations. The daytime $Z_h$ was reasonably estimated by the reanalysis and the parameterization schemes. However, the nighttime $Z_h$ is significantly overestimated by the ERA-I and parameterization schemes, along with a poor representation of the transition of $Z_h$ at dusk and dawn. The estimated $Z_h$ was poorly correlated at night with observations and the spatial variability of $Z_h$ over the CONUS is not well represented. In general, a deeper $Z_h$ corresponds to mixing over a deeper column and a high near-surface wind speed. Similarly, GFS and NAM provide a reasonable daytime near-surface wind speed and an overestimated nocturnal nighttime wind speed, which results in the lowest correlation with observations around sunset, which in turn results in a maximum improvement in RC by the postprocessing methods at sunset.

All the postprocessing methods reduce the bias of the 10-m wind speed prediction from NAM (Fig. 10a), although not equally. All methods effectively reduce the bias in the GFS forecasts to a statistically equal degree (Fig. 10b). The analog methods offer relatively better performance compared to 7-Day and KF for NAM and RUC predictions (Figs. 10a,c).

The greatest reduction in CRMSE from postprocessing occurs for minima in NWP CRMSE, but the greatest reduction in bias occurs for maxima in NWP bias. For GFS, the effectiveness of the bias reduction is clear from Fig. 10b. More subtly, comparing the NAM CRMSE difference (Fig. 9a) with the bias difference (Fig. 10a) during the night shows that the maximum reduction of
random errors coincides with the minimum reduction in systematic error. As noted above, the maximum difference in CRMSE occurs during the night. We speculate that the maxima in NAM’s daytime CRMSE are difficult to reduce. Although the NAM bias shows only a weak diurnal variability (Fig. 10a), the maxima are also in the daytime, corresponding to the maxima in bias reduction from postprocessing.

2) AIR TEMPERATURE AND RELATIVE HUMIDITY

Postprocessing reduces the CRMSE and increases the RC of T for all forecast lead times (Figs. 11a,b). Recall that for wind speed, all the methods exhibit a prominent diurnal variation in the difference in CRMSE. The maximum reduction in CRMSE is at night with a peak in RC that decreases over the course of the night (Fig. 11b). Different from the wind speed results, both the CRMSE and RC improvements coincide with high CRMSE values at night for the NAM forecasts (Figs. 11a,b). Figure 11b shows that the magnitudes of the differences in RC are relatively small. This is because NAM’s RC for T forecasts is already high (between 0.87 and 0.96; not shown). Temperature biases, already less than 0.5° C in NAM, are easily reduced (cf. Figs. 11b and 10d) by the postprocessing; the six maxima and minima correspond to the six maximum-magnitude biases in NAM.

Why the CRMSE is most reduced at the same time as CRMSE from the NWP is largest can be explained in terms of error time scales. The postprocessing methods are not designed to explicitly minimize either CRMSE or bias, as calculated over the 2-month verification period. For a given forecast hour, the bias in the verification is simply the 2-month mean error, and the CRMSE is the 2-month standard deviation of the error. 7-Day addresses errors at that time scale, and KF addresses errors at some shorter time scales. AN and ANKF do not have any relationship to a particular time scale. All postprocessing methods have the potential to reduce both the bias and the CRMSE. In this study, the CRMSE of T over the 2-month period is measuring T errors that evolve faster than in 2 months, but slower than the errors in wind evolve.

The CRMSE and RC differences for RH (Figs. 11c,d) exhibit a pronounced diurnal variation for predictions by the analog methods. AN and ANKF perform similarly, and both improvements are significantly different from zero. KF and 7-Day also perform similarly to each other, and with less diurnal variability than from the analog methods. The diurnal variability in the analog methods may arise because T is one of the predictor variables in RH prediction, but is not a predictor in KF and 7-Day error predictions. Figure 10e shows that RH biases can be reduced when they are large (Fig. 11d), as in the case of temperature, with 7-Day and KF somewhat overcorrecting for the nighttime dry bias. Overall, the analog methods predict a more accurate RH
compared to 7-Day and KF, and the results can be understood by invoking the same error-time-scale arguments put forth for temperature.

Results from this section are broadly consistent with the previous sections, but add to the picture by showing a strong diurnal dependence of the performance of the postprocessing methods. The diurnal variability in each metric can be linked, in most cases, to the diurnal variability of random and systematic errors in the NWP forecasts prior to postprocessing. Because random and systematic errors can also vary in space, we next evaluate how postprocessing varies across the model domain.

3) CRMSE and RC Variability in Space

The relative magnitudes of systematic and random errors can vary according to season and the regional climate within that season. For example, random errors may contribute more to the overall error where thunderstorms are common. Differences in CRMSE and differences in RC at each station are computed by averaging over the evaluation period and over all the lead times. The diurnal variability shown in the previous section will lead to some cancellation, but some spatial variability remains. Here, we focus only on results that add further insight and avoid an exhaustive report. In each plot, colored circles represent the value of the statistic (CRMSE difference or RC difference) at the location of each station, with warmer colors (positive values) indicating improved forecast accuracy (i.e., smaller CRMSE and higher RC produced by the postprocessing method).

Consistent with results reported earlier, Fig. 12 shows that the AN method is most effective at reducing wind speed CRMSE compared to raw GFS forecasts. Although it appears that 7-Day is not an improvement, Fig. 9b showed that on average it does offer an improvement of less than 0.5 m s\(^{-1}\). KF, ANKF, and AN appear to improve CRMSE progressively more, which is again consistent with earlier results.

For GFS, the spatial variation across the CONUS is weak, but all postprocessing methods result in greater improvement over coastal California. Although we cannot be sure at this point, we speculate that this reveals persistent effects of the coastline, such as through sea breezes.

It is apparent that the analog methods (especially AN) yield sizeable reductions in CRMSE near 100°W, a region of high topography where the predictive skill of the forecast is lower because of the challenge of modeling atmospheric flow over complex terrain. A similar finding was reported by Delle Monache et al. (2011) in the same region for predictions by WRF.

AN and ANKF show slightly less improvement at a few stations over the northern plains, indicated by a swath of yellow among the orange (Fig. 12). This is
where we expect mesoscale convective systems to be active in June and July. KF reduces the CRMSE more than 7-Day, because KF addresses errors at shorter time scales (Fig. 12, top). On the other hand, a significantly improved performance by AN and ANKF over the southeastern CONUS is evident (Fig. 12, bottom). Laing and Fritsch (1997) showed that numerous deep convective events occur over the southeastern CONUS, with the plains dominated by few MCS events. This is also reflected in the relatively numerous lightning events (Houze 2004; Holle et al. 2011) over the southeastern CONUS compared to the plains. This suggests that the availability of numerous events (or good quality analogs) in the training data over the southeastern CONUS is likely to result in a relatively better performance by AN and ANKF compared to the northern plains. Results are similar for CRMSE using NAM and RUC (not shown).

Figure 13 shows that all the methods also improve the GFS RC broadly over the CONUS. Relatively large increases are again evident along the coast of California and, additionally, along the Gulf of Mexico. Compared to other methods, AN improves upon the raw RC near 36°N, 100°W, and the northeastern CONUS regions of complex terrain and an area of mesoscale convective system activity, respectively. Thus, over the CONUS and for GFS, the most accurate forecast of wind speed is from AN.

In the case of RUC, the RC improves progressively between KF and AN, but the greatest improvement is produced by ANKF (Fig. 14). While the RC is improved by AN along the West Coast and the Gulf of Mexico, ANKF provides improvement over much of the CONUS. This suggests that for the RUC, the errors associated with the best analogs may be a good predictor of the current forecast error in many parts of the CONUS; the RUC has systematic errors that are strongly state dependent rather than time dependent. A similar result from RC for NAM (not shown) is found, consistent with the findings for the WRF Model by Delle Monache et al. (2011). Because of the small RC in the GFS forecasts, as noted earlier, the magnitude of the improvement is much greater than possible for the RUC and NAM, for which RC is already high. As also noted earlier, we expect this is simply a result of the higher resolution in the RUC and NAM, which resolve faster time scales and allow for variability closer to the observed variability.
KF, ANKF, and AN all improve the CRMSE for the NAM RH forecasts (Fig. 15). The 7-Day also makes a smaller improvement (1.0%–2.0% in Fig. 11c), but the CRMSE is made worse, on average, in some regions. The northern Great Plains, where thunderstorms are active during June and July, are not amenable to correction by the slow time scale error estimates from 7-Day. As before, results for RH RC are overall quite similar and are not shown.

Overall, these results show that the analog-based methods, and to a lesser extent KF, can improve CRMSE and RC broadly over the CONUS. Given sufficient data, these methods appear reasonably resistant to variations in climate and the associated error components.
Although 7-Day shows some benefit on average, in some regions this method of postprocessing can be harmful to the random error component, even if it reduces bias.

d. Seasonal variation

The performance of the postprocessing methods as a function of seasons is studied by repeating the exercise of the previous sections for different verification periods. Earlier, the verification period covered June and July 2011. Five additional 2-month verification periods are used: August–September, October–November, December–January, February–March, and April–May. As before, the training datasets are 10 and 12 months long for the first and last forecast verification day, respectively. In general, the results appearing in the previous section continue to hold with few exceptions, which are highlighted next.

1) Wind Speed

The forecast error of the raw model can vary as a function of seasons. The errors characterized by bias and CRMSE for GFS do not show significant variation with season. On the other hand, the maximum and minimum errors occur at different times of the year for NAM and RUC. The smallest (largest) errors for NAM and RUC occur during June–July (April–May).

The MAE viewed as a function of season shows that AN and ANKF remain the best postprocessing methods and exhibit a similar behavior (Figs. 3 and 4) in forecast improvements for all models (not shown).

A postprocessed forecast having zero CRMSE is ideal and will result when the difference (raw minus postprocessed) CRMSE equals the raw CRMSE. The difference CRMSE (e.g., Fig. 9a) for GFS shows no significant variation with season (not shown). This suggests a season-invariant performance of AN and ANKF for GFS. The RUC forecasts are most improved during April–May when the CRMSE are the largest (not shown). For the NAM, the most improvement occurs during December–January (Fig. 16a). Thus, AN and ANKF reduce the raw CRMSE of all the models throughout the year.

The spatial distribution of the difference CRMSEs for all models vary with season. However, AN and ANKF show the largest improvements in CRMSE compared to 7-Day and KF throughout the year for all the models (not shown). The least improvement for NAM occurs in June–July (Fig. 3b) with the most CRMSE improvement occurring during December–January (Fig. 17). Thus, the analog methods reduce CRMSE the most and do so throughout the year.

2) Temperature and Relative Humidity

The bias and CRMSE of raw NAM temperature and relative humidity forecasts exhibit a minimum (maximum) between June and July (December and January). The NAM model underwent a significant upgrade during March 2011, which significantly improved the temperature forecasts (information online at http://www.emc.ncep.noaa.gov/mmb/briefings/HiResW-upgrade.)
The analog method works best when analogs come from a training dataset produced by a frozen version of the forecast model. The use of a training dataset coming from two different versions of the NAM model is likely to result in picking less representative analogs of the current forecast impacting AN and ANKF.

AN and ANKF improve the MAE better than 7-Day and KF between June and July (Fig. 6a). However, during the rest of the year the improvement by AN and ANKF is less than that of 7-Day and KF. The former result arises because out of the total 10–12 months in the training dataset, 7 months of NAM forecasts came from a model version [WRF–Nonhydrostatic Mesoscale Model (NMM), version 2.2] prior to its upgrade (to version 3.2). The latter result is likely due to 5 months of the forecasts coming from version 2.2. A typical performance of AN and ANKF between October and March, when the training dataset contains about 5 months of NAM version 2.2 forecasts, is shown in Fig. 16. Comparing Figs. 6a and 16b demonstrates the impact on the AN and ANKF results as a result of mixing forecasts from two different versions of NAM. Given this limitation of the temperature training record from NAM, the reduction in CRMSE as a function of season is not addressed.

Based on MAE improvement, AN and ANKF outperform 7-Day and KF for relative humidity for all months (not shown) with the largest improvement seen during October–November (Fig. 16d). The CRMSE of RH is improved at all lead times by AN and ANKF and is significantly better during December and January (Fig. 16c) compared to June and July (Fig. 11c). The spatial distribution of the difference CRMSE of RH varies with season and also reveals a maximum reduction in CRMSE by AN and ANKF during October–November (not shown).

e. Operational model output statistics

We compare the postprocessing methods as a function of lead time against the operational MOS. The NAM model was upgraded in March 2011, which significantly impacted the temperature forecasts while having no significant impacts on the wind speed and dewpoint forecasts. The moisture information from the NAM MOS is available in the form of dewpoint temperature $T_d$. A consistent MOS relative humidity can be derived from $T_d$ by using conversion formulas from the operational NAM model used in 2011. The period between 2010 and 2011 saw a transition of the NAM model from the Nonhydrostatic Mesoscale Model to NMM-B where B refers to an Arakawa B grid with the transition being completed in October 2011 (http://www.nws.noaa.gov/os/notification/tin11-16nam_changes.txt). This renders difficult a conversion of $T_d$ to RH in a consistent way and precluded the comparison of the MOS RH.

Since the upgrade significantly impacted the NAM $T$ predictions (section 3d), with the result being a training data that came from different versions of the NAM...
model, we have excluded the MOS T predictions in our analysis. Thus, this section includes MOS wind speed from NAM and GFS. Further, a lack of wind speed MOS data from RUC also precluded its comparison against the other postprocessing methods. The comparison is performed for June and July 2011.

Recall that we use a total of 740 stations and a 72-h lead time for GFS and NAM. The MOS data are missing at about 80 stations and at several lead times (0, 3, 63, and 69 h). Further, the MOS are produced using a much longer training time period (2–3 yr) using a larger set of predictor variables compared to the postprocessing methods. Thus, the results of the following section should be interpreted with great care and a thorough attribution of performance differences between MOS and the other methods is therefore not attempted.

**WIND SPEED**

The CRMSE and RC as a function of lead time are shown in Figs. 18a and 18b, and in Figs. 18d and 18e, respectively. The CRMSE performance of MOS from NAM and GFS are comparable to ANKF at most of the lead times. However, AN outperforms MOS. A degraded performance in CRMSE from MOS is evident at certain lead times (15 and 24, and 39 and 45 h). The degraded performance for NAM is pronounced. One reason for this pronounced degradation may be that the NAM model underwent a major upgrade, which resulted in the use of a hybrid training dataset coming from two different versions of the forecast model. Nevertheless, the degraded performance between certain lead times in both GFS and NAM is speculated as arising because of the models struggling with the prediction of the evening time PBL transition phase. While no significant difference is evident between AN and MOS for the RC from NAM, the MOS RC from GFS significantly outperforms the other postprocessing methods (Figs. 18d,e).

On the other hand, during periods of degraded performance in CRMSE (Fig. 18a), the NAM bias (Fig. 18c) shows a pronounced improvement while outperforming the other postprocessing methods. However, at the other lead times, the postprocessing methods outperform the MOS. There is no significant difference between the performance of the bias from GFS MOS compared to the other methods (Fig. 18f). Thus, the analog methods are competitive compared to the MOS.
while outperforming the MOS in reducing the random errors during certain lead times.

4. Summary and conclusions

Delle Monache et al. (2011) proposed two new analog-based postprocessing methods to increase forecast accuracy. The purely analog-based method (AN) uses a weighted average of the observations associated with the best analogs to produce a postprocessed forecast. Alternately, a Kalman filter (KF) predictor–corrector algorithm can be applied to the analogs, ordered according to their fit to the current forecasts, to get an exponentially weighted estimate of the error associated with the analogs. That error is then used to postprocess the current forecast; this is the ANKF method. KF can also be applied directly to the time series of past forecasts. Finally, a simple 7-day running-mean error (7-Day) can correct the forecast. Herein, AN and ANKF were evaluated against the time series–based KF and 7-Day methods, over the CONUS, for near-surface winds, temperature $T$, and relative humidity (RH) predicted by GFS, NAM, and RUC. The objectives were to test each method’s performance and to gain greater insight into their strengths and weaknesses.

The relative levels of random and systematic error in the raw NWP forecasts determine, to a large extent, how well each postprocessing method reduces forecast errors. The NAM and RUC, at much finer spatial resolution than the GFS, have greater randomlike errors simply because they simulate faster time scales. Forecasts for 2-m $T$ have a relatively greater systematic error component, while 2-m RH and 10-m winds have a relatively greater random component.

When random errors dominate, analog-based methods far outperform time series–based methods. When systematic errors dominate (i.e., a bias), analog methods lose their advantage over the time series methods. Systematic errors evolve on slow time scales and are easy to estimate. The relatively short 10–12-month training dataset used in this study may be one of the causes of the residual bias left in the analog-based estimates. In particular, the AN forecast that is based on past observations would provide an unbiased prediction with an infinite training data record length. Thus, by using a multiyear historical dataset that is often available at operational centers, it is anticipated that the bias may be further reduced.

For 10-m wind speed, AN and ANKF led to positive improvements when day-to-day changes in error ($|\text{MAE}_d - \text{MAE}_{d-1}|$) varied between 0.25 and $8 \text{ m s}^{-1}$, with AN offering the greatest improvement over the raw GFS, except for when the error change is very small. Very small error changes suggest persistence in the errors, which are easily handled by 7-Day and KF. The maximum improvement by any method over the raw NAM reached about 30%, with ANKF, KF, and 7-Day showing slightly lower MAE than AN for small variations in $|\text{MAE}_d - \text{MAE}_{d-1}|$ (between 0.25 and $1.25 \text{ m s}^{-1}$). The exponential weighting strategy used by ANKF and
KF led to greater forecast improvements for small day-to-day error changes. However, when the error was larger (>1.25 m s\(^{-1}\)), AN was the best method because AN uses weights that are directly proportional to how close the analogs are to the forecast. Results based on RUC were similar to those based on NAM.

From all three models, wind speed CRMSE was consistently reduced and RC consistently increased across the CONUS, with AN and ANKF performing best. Details among the methods varied. The 7-Day showed some areas where the CRMSE and RC became worse, such as in the northern plains, where thunderstorms are common in June and July.

Analysis for NAM T and RH showed qualitatively similar results, but highlighted that T errors evolve more slowly than do wind speed errors. The AN showed a smaller (compared to 7-Day, KF, and ANKF) improvement for \(|MAE_d - MAE_{d-1}|\) T and RH values less than 3°C and 10%, respectively, with larger improvements for greater day-to-day error changes.

All postprocessing methods exhibited a nocturnal maximum in the reduction of CRMSE from all the models. This was attributed to the poor representation of the nighttime PBL processes in the models resulting in a poor correlation of forecasts with observations. The different methods produced less distinctly different results for the GFS, apparently because it is dominated by slower-error processes. We speculate that the relatively smaller daytime reduction in the CRMSE results from randomlike PBL processes, clouds, and deep convection. AN and ANKF consistently produced the smallest CRMSE at all forecast lead times for all three models.

Improvements in CRMSE and RC by forecast lead time differed little between the two analog methods. RC improvement was small for the NAM and RUC because their NWP forecasts were already highly correlated with observations. RH and T behaved similarly, with the analog predictions exhibiting a prominent diurnal variation compared to 7-Day and KF. This is attributed to the use of T as a predictor in the analog methods.

Plots of bias against CRMSE showed that for the 10-m wind speed, T, and RH, the analog methods reduced the CRMSE the most. The 7-Day and KF were effective at nearly eliminating the bias. In general, the reduction in CRMSE outweighed the reduction in bias so that the analog methods were more effective at reducing the RMSE. Results from NAM and RUC were similar; bias correction for the GFS offered relatively more benefit because it lacks the fast scales that impact CRMSE.

Our results indicate that the direct analog method (AN) is most effective at improving forecast skill as measured by CRMSE and RC and consequently RMSE across a broad range of lead times, locations, error levels, and day-to-day error variability. The variation of the results as a function of the season was studied by repeating the exercise for a whole year. Overall, the conclusions reached for June and
July were confirmed to be valid for the rest of the year for all models and variables with the caveat that the NAM T prediction significantly improved around March 2011 as a result of a major model upgrade.

The analog methods were compared against operational MOS for wind speed (GFS and NAM) and temperature (NAM). The results show that the analog methods are competitive compared to the MOS while outperforming the MOS in reducing the random errors at certain lead times.

Acknowledgments. This research was supported by the U.S. Army Test and Evaluation Command and Defense Threat Reduction Agency through an interagency agreement with the National Science Foundation. We thank the editor and the two anonymous reviewers for their constructive suggestions that significantly improved the manuscript.

APPENDIX

Kalman Filter (KF) Method

In the Kalman filter (KF) method, past prediction errors at a given location (i.e., the difference between forecasts and observations) are used to estimate the bias in the current raw forecast.

The true (unknown) forecast bias $x_i$ at time $t$ is modeled by the previous true bias plus a noise term $\eta_i$, which is normally distributed with zero mean and a variance $\sigma^2_{\eta,i}$ (Bozic 1994):

$$x_i = x_{i-\Delta t} + \eta_i,$$  \hspace{1cm} (A1)

where $\Delta t$ is a time lag. The KF approach further assumes that the forecast error $y_i$ (forecast minus observation at time $t$) contains an unsystematic error term $e_i$, where $e_i$ is normally distributed with zero mean and variance $\sigma^2_{e,i}$. This arises because of unresolved terrain features, numerical noise, lack of accuracy in the physical parameterizations, and errors in the observations:

$$y_i = x_i + e_i = x_{i-\Delta t} + \eta_i + e_i.$$  \hspace{1cm} (A2)

In the KF method, following the Kalman filter methodology (Kalman 1960), the recursive predictor of $x_i$ (derived by minimizing the expected mean-square error) is expressed as a combination of the previous predicted bias and the previous forecast error:

$$\hat{x}_{i+\Delta t|t} = \hat{x}_{i|\Delta t} + K_i (y_i - \hat{x}_{i|\Delta t}),$$  \hspace{1cm} (A3)

where the hat (’’) indicates the estimate. The weighting factor $K_i$, called the Kalman gain, is computed from

$$K_i = \frac{p_{1-\Delta t} + \sigma^2_{\eta,i}}{(p_{1-\Delta t} + \sigma^2_{\eta,i} + \sigma^2_{e,i})},$$  \hspace{1cm} (A4)

where $p$ is the expected mean-square error and is computed as

$$p_t = (p_{1-\Delta t} + \sigma^2_{\eta,i})(1 - K_i).$$  \hspace{1cm} (A5)

In contrast to the 7-day running mean, it can be seen as an exponential weighted mean where recent errors are weighted more than past errors.

It can be shown (Dempster et al. 1977) that the time series

$$z_i = y_{i+\Delta t} - y_i = \eta_i + e_{i+\Delta t} - e_i$$  \hspace{1cm} (A6)

has variance

$$\sigma^2_z = \sigma^2_{\eta} + 2\sigma^2_e.$$  \hspace{1cm} (A7)

Assuming

$$r = \sigma^2_{\eta}/\sigma^2_e,$$  \hspace{1cm} (A8)

Eq. (A7) becomes

$$\sigma^2_z = r\sigma^2_{\eta} + 2\sigma^2_e = (2 + r)\sigma^2_e.$$  \hspace{1cm} (A9)

The KF algorithm will quickly converge for any reasonable initial estimates of $p_0$ and $K_0$. The filter algorithm is run independently on data for each forecast lead time, using values from previous days at the same forecast lead time hour. Thus, the KF approach adapts its coefficients as new forecasts and observations become available.

The variance $\sigma^2_{\eta,i}$ is estimated in three steps. First, the expected mean-square error $\sigma^2_{e}$ is estimated by applying Eq. (A5). Second, the new Kalman gain is computed using Eq. (A4). Finally, $\sigma^2_{\eta,i}$ is estimated by using Eqs. (A3) and (A8).

REFERENCES


