Influence of Surface Observations in Mesoscale Data Assimilation Using an Ensemble Kalman Filter

SO-YOUNG HA AND CHRIS SNYDER
National Center for Atmospheric Research, Boulder, Colorado

(Manuscript received 6 April 2013, in final form 22 November 2013)

ABSTRACT

The assimilation of surface observations using an ensemble Kalman filter (EnKF) approach was successfully performed in the Advanced Research version of the Weather Research and Forecasting Model (WRF) coupled with the Data Assimilation Research Testbed (DART) system. The mesoscale cycling experiment for the continuous ensemble data assimilation was verified against independent surface mesonet observations and demonstrated the positive impact on short-range forecasts over the contiguous U.S. (CONUS) domain throughout the month-long period of June 2008. The EnKF assimilation of surface observations was found useful for systematically improving the simulation of the depth and the structure of the planetary boundary layer (PBL) and the reduction of surface bias errors. These benefits were extended above PBL and resulted in a better precipitation forecast for up to 12 h. With the careful specification of observation errors, not only the reliability of the ensemble system but also the quality of the following forecast was improved, especially in moisture. In this retrospective case study of a squall line, assimilation of surface observations produced analysis increments consistent with the structure and dynamics of the boundary layer. As a result, it enhanced the horizontal gradient of temperature and moisture across the frontal system to provide a favorable condition for the convective initiation and the following heavy rainfall prediction in the Oklahoma Panhandle. Even with the assimilation of upper-level observations, the analysis without the assimilation of surface observations simulated a surface cold front that was much weaker and slower than observed.

1. Introduction

Compared to other routine in situ observations, surface observations (2-m temperature and moisture, 10-m winds) have relatively good spatial and temporal resolution, but their use in atmospheric data assimilation has proven substantially more difficult. Indeed, many state-of-the-art systems do not assimilate surface observations directly; at the European Centre for Medium-Range Weather Forecasts (ECMWF), for example, surface observations are used in a two-dimensional surface analysis that forces the land surface model but are not used in the variational assimilation scheme for the atmosphere (de Rosnay et al. 2013). The difficulty of assimilating surface observations arises from at least two issues. First, the assimilation scheme must specify the background error covariances, especially in the vertical and in relation to the structure of the planetary boundary layer (PBL); the background error covariances are important because, in statistical data assimilation schemes, they specify how surface observations will influence the analysis. Second, forecast models typically have limited skill in the PBL owing to significant errors in the parameterizations of vertical mixing, surface fluxes, and land surface processes. This paper concentrates on the first issue and presents results showing the effectiveness of the ensemble Kalman filter (EnKF) for the assimilation of surface observations, which we attribute to the EnKF’s use of background error covariances computed from an ensemble of short-range forecasts and that are therefore consistent with the local structure and dynamics of the PBL.

Specifying the background error covariances in the PBL is challenging because of complexity and variability in the structure of the PBL and in the relation of surface variables to the rest of the atmosphere. Physical intuition suggests that observed surface variables may often have strong correlations with variables within the PBL and weak correlations with flow outside the PBL, depending on the time of day and the boundary layer...
regime. Nudging schemes for assimilation have incorporated this intuition by requiring the nudging coefficients for surface observations to decrease to zero above the PBL top (Stauffer et al. 1991). Such modifications to the coefficients are, however, heuristic, in that they are based on the developer’s assumptions for the relation of surface variables to other variables and other levels.

Three-dimensional variational data assimilation (3DVAR) schemes typically use stationary covariance, calculated from statistics of forecast differences, which can capture basic aspects of PBL dynamics such as the correlation between vertical vorticity and horizontal divergence at synoptic scales (Derber and Bouttier 1999). Nevertheless, because they use stationary background covariances, 3DVAR schemes generally do not account for any dependence of analysis increments on the instantaneous structure, the diurnal cycle, or the regime of the PBL (e.g., Benjamin et al. 2004).

More sophisticated statistical assimilation approaches, such as four-dimensional variational data assimilation (4DVAR) and the EnKF schemes, extract dynamical information directly from the forecast model and produce analysis increments that depend on the local state of the PBL and its recent history. In the EnKF, the relative weighting in the analysis between the background and observations, and the influence of observations away from the station locations, are determined by anisotropic and inhomogeneous background covariances. These covariances, in turn, are estimated from an ensemble of short-range forecasts using the full, nonlinear model, and the covariances will, therefore, change as the PBL evolves.

Several previous studies have specifically examined the assimilation of surface observations with an EnKF method. Hacker and Snyder (2005) considered simulated surface observations in a single-column model. Their results confirmed the physical intuition that surface observations are strongly correlated with, and therefore when assimilated informative of, conditions throughout the well-mixed PBL or even into the previous day’s residual layer. Fujita et al. (2007) assimilated actual surface observations in the EnKF. In addition to examining the role of forecast model error, they demonstrated improvements to the analysis of important mesoscale features for two severe weather cases as well as improvements lasting 6–12 h in the subsequent forecasts. Meng and Zhang (2008) also briefly examined assimilation of surface observations in a case study of a mesoscale convective vortex but found little influence on the short-range forecast. Results from a longer, 5-week period of assimilation appeared in Ancell et al. (2011). They did not examine the influence of surface observations on the analysis directly, but did show that special treatment of surface observations, either through bias correction of the prior forecasts or by reducing their observation-error variance, resulted in better fits of the analysis and 6-h prior forecasts to unassimilated surface observations. These benefits did not, however, extend above the surface.

In the present study, we continue the exploration of assimilation of real surface observations in the EnKF approach using the Advanced Research version of the Weather Research and Forecasting Model (WRF) coupled with the Data Assimilation Research Testbed (DART) system (WRF–DART) and examine the influence of surface observations on mesoscale analyses in an optimized environment over the contiguous U.S. (CONUS) domain. Our goal is to investigate the effectiveness of the EnKF in the assimilation of surface observations and demonstrate improvements of the following forecasts in the boundary layer and farther aloft, as well as at the surface. For this purpose, we perform experiments with and without assimilation of surface observations, and make forecasts from the mean EnKF analysis for 72 h. Through a mesoscale convective case study, we examine how the assimilation of surface data can provide a favorable condition for the initiation and the development of severe convection associated with a squall line and to what extent the analysis increment affects the structure and evolution of the boundary layer.

Surface pressure observations reflect the distribution of mass vertically integrated through the depth of the entire atmosphere, and their use in the EnKF has been explored by Dirren et al. (2007) and Compo et al. (2006), for example. It will not be a main focus here, since they are not particularly sensitive to or informative of conditions in the PBL and their assimilation is subject to neither of the two main issues noted in the first part of this session.

The data assimilation system implemented here largely ignores forecast model errors, despite their potential importance in the PBL and therefore in the assimilation of surface observations. The system includes only a spatially and temporally varying, adaptive inflation (Anderson 2009), which compensates for underestimation of the analysis variance by the EnKF as well as underdispersion in the prior ensemble forecasts owing to unrepresented model error. Other possible approaches aimed specifically at representing model uncertainties within the EnKF are employing multiple suites of physical parameterizations (Fujita et al. 2007; Meng and Zhang 2008; Hacker et al. 2011) and including an additive-noise or “backscatter” scheme within the model (Hacker et al. 2011; Berner et al. 2011). We will examine these approaches in a subsequent study.
The outline of the paper is as follows. The experimental design is outlined in section 2, including our configuration of the Advanced Research version of WRF (Skamarock et al. 2005) and details of the EnKF as implemented in the DART (Anderson et al. 2009). Section 3 covers the observations used in the assimilation and the observation processing. In particular, we describe in detail how the error variances for surface observations were assigned because of their importance to our results. Results from the month-long assimilation experiments with and without surface observations appear in section 4, followed by a case study of how assimilation of surface observations affects the forecast of severe convection that occurred during the assimilation cycling period. A summary and conclusions are provided in section 6.

2. Experiments

Our data assimilation experiments were performed in a full cycling mode for a 1-month period of June 2008, starting from 0000 UTC 1 June, at a 3-h interval. That is, observations were assimilated every 3 h using the 3-h forecast ensemble from the previous cycle to compute the prior (or background) ensemble covariance, and producing the analysis ensemble as an ensemble of initial conditions for the next 3-h forecasts. Both the analysis and forecast steps used a 50-member ensemble in all experiments.

a. The model configuration

The forecast step employed WRF version 3.2 on the domain depicted in Fig. 1, with a horizontal grid of 123 × 99 points at 45-km resolution for the outer domain (D1) and of 163 × 106 points at 15-km for the inner domain (D2). The 3-h ensemble forecasts in the two domains ran with two-way nesting. We used 41 vertical levels to discretize the atmosphere between the surface and 50 hPa, with about 10 layers in the lowest 2 km. On both domains, the model physics configuration included the Yonsei (YSU) PBL scheme (Hong et al. 2006), the Kain–Fritsch (Kain 2004) cumulus parameterization, the WRF single-moment 5-class microphysics scheme (WSM5; Hong et al. 2004), the Noah land surface model (LSM; Chen and Dudhia 2001), Dudhia shortwave (Dudhia 1989), and Rapid Radiative Transfer Model (RRTM) longwave (Mlawer et al. 1997) radiation schemes. In the current WRF version, each PBL scheme is tied to a particular surface-layer scheme, and the YSU PBL uses the fifth-generation Pennsylvania State University–National Center for Atmospheric Research (PSU–NCAR) Mesoscale Model (MM5) surface-layer similarity
(Zhang and Anthes 1982). Here, all the ensemble members use the same model configuration.

The ensemble-mean lateral boundary conditions (LBCs) on D1 were given by the $1^\circ \times 1^\circ$ National Centers for Environmental Prediction (NCEP) Global Forecast System (GFS) 6-h forecast valid at the appropriate times, consistent with what would be available were the system running in real time. We accounted for the uncertainty in the LBCs, following Torn et al. (2006), by specifying each member’s LBCs as the sum of the ensemble-mean LBCs and random, spatially correlated perturbations. The perturbations were generated as draws from a Gaussian random vector, using the background covariance with random $\text{cv} = 3$ in the WRF data assimilation system (Huang et al. 2009), and variances, and horizontal and vertical length scales tuned as in Torn and Hakim (2008).

At the beginning of the analysis cycle, initial conditions are required for each member. On D1, these were generated in the same way as the ensemble of LBCs, but centered on the GFS analysis from 0000 UTC 1 June 2008. On the inner domain, there were no initial perturbations; these spin up over 1 or 2 days through the interaction with the outer domain. The initial soil states for each member are simply the GFS analysis, without additional perturbations. The details of the ensemble initial conditions at 0000 UTC 1 June have little effect on our results, as the assimilation of observations and random perturbations in the LBCs cause the ensemble to become largely independent of initial conditions within a few days.

b. The EnKF configuration

For assimilation, we chose the EnKF, implemented in DART as an ensemble adjustment Kalman filter (Anderson 2001), one variant in the class of deterministic “square root” EnKFs (Tippett et al. 2003). The analysis variables were the three components of velocity, potential temperature, column mass of dry air, geopotential and mixing ratios for water vapor and microphysical species, as well as various surface diagnostic variables available in WRF (as specified in the following paragraph). Soil states were not updated by the assimilation system and therefore evolve over the month of cycling according to the atmospheric forcing from each member.

Surface observations assimilated were 2-m temperature and dewpoint (T2 and Td2, respectively), the 10-m horizontal wind components (U10 and V10), and the surface alimeter. The forward operators for these observations were computed as follows. At the end of the forecast step, we output gridded fields of 2-m water vapor mixing ratio, T2, U10, V10, and surface pressure for each member, which were diagnosed in the PBL and surface-layer schemes as part of the normal WRF integration. Within the EnKF, these quantities were first used, if necessary, to calculate observed variables (in the case of Td2 and alimeter setting), and then interpolated horizontally to the observation location to provide a prior (or background) estimate of the observations from each member. These forward operators are thus consistent with the PBL and surface-layer schemes in the forecast model.

Fujita et al. (2007) followed a similar approach, but diagnosed T2, Td2, U10, and V10 within the EnKF using surface-layer similarity relations. Note also that these forward operators ignored differences between model terrain height and station elevations, unlike other studies where T2 was corrected based on the local lapse rate (Benjamin et al. 2004).

Besides the check on the station elevation to ensure that it is close enough to the model’s orography (which will be described in section 3), the ensemble data assimilation system also discards observations based on the innovation magnitude (a so-called “outlier check”). An observation is rejected if the absolute value of its difference from the mean prior observation is greater than 3 times the total spread, which is given by the square root of the sum of the prior ensemble variance and the observation-error variance.

To counteract the detrimental effects of sampling error, we utilized a distance-dependent covariance localization (Houtekamer and Mitchell 2001; Hamill et al. 2001). For the algorithm used here, in which observations are processed serially within the EnKF, covariance localization amounts to multiplying the observation-state covariance estimated from the ensemble by a function of the distance between the observation and state variables, which decreases smoothly to zero at a finite distance according to the compactly supported correlation function of Gaspari and Cohn (1999). Based on limited sensitivity tests for a two-week period (not shown), we took that distance to be 600 km horizontally and 8 km vertically.

Even using the covariance localization, however, other factors contribute to the underestimation of analysis uncertainties in our system, including deficiencies in the forecast model (which we ignored in running ensemble forecasts), effects of nonlinearity and non-Gaussianity, and remaining sampling errors. The influence of these will vary spatially and temporally depending on the details of flow and of the observation network. To account for the combined effects of these factors, we employed spatially and temporally varying, adaptive inflation in the model state space (Anderson 2009).
The WRF-DART system updates multiple, nested domains in a straightforward fashion. First, the forward operators are computed using the fields from the finest grid that contains the observation. Then, the analysis proceeds on all domains simultaneously, with each observation updating all grid points within the localization radius, regardless of the domain to which they belong. Thus, observations within the inner 15-km domain can also affect the analysis on the outer domain, and vice versa.

Since surface observations are typically reported at least every hour, the frequency of assimilation is a potentially important question. Our choice of a 3-hourly assimilation cycle was based on initial tests over the first 10-day period of cycles that compared the performance of 1-, 3-, and 6-hourly cycling frequencies and revealed that innovations for surface observations were not very sensitive to the frequency of assimilation (not shown). Choosing 3-hourly cycles was a compromise between increasing the number of surface observations and the indications from some previous studies that more frequent assimilation cycles lead to spurious noise in the background forecast (Benjamin et al. 2004).

3. Observations

In this study, we collected observations from the Meteorological Assimilation Data Ingest System (MADIS) of the National Oceanic and Atmospheric Administration (NOAA) for different observation types. Radiosonde soundings, Aircraft Communications and Reporting System (ACARS), marine, and METAR data were used for assimilation. We also collected surface observations from various mesonet archived in MADIS. Because they have slightly lower quality than the METAR observations but show better coverage over the CONUS domain, we withheld the mesonet observations from the assimilation and used them as independent observations to evaluate the quality of the surface analyses and forecasts. The mesonet observations were otherwise treated identically to the METAR observations: the same forward operators, the same observation-error variances, and the same quality control were applied. The ACARS observations were assimilated as a fewer set of superobservations, given by averages over boxes 45 km on a side and 50 hPa deep.

Specification of observation-error variances is an important element of the successful data assimilation of real observations. For all observations except surface data, we employed observation-error variances taken from the gridpoint statistical interpolation (GSI) system (Kleist et al. 2009). For surface observations, we have conducted additional tuning based on observation-space diagnostics.

Although more sophisticated approaches are available (Desroziers and Ivanov 2001; Desroziers et al. 2005), we tuned the observation-error variances using the fact that, when observation and background errors are unbiased and uncorrelated,

$$\text{tr}((d d^T)) = \text{tr}(P) + \text{tr}(R),$$

where $d = y - \langle H(x) \rangle$ is an innovation, $y$ is the observation vector, $x$ is the state vector, $H$ is the forward operator, $P = \text{cov}[H(x)]$ is the background covariance matrix for the observed variables, $R$ is the observation-error covariance matrix, and angle brackets denote an expectation over both the observation-error and the background distribution. This equation is routinely used in both ensemble forecasting and ensemble data assimilation, with the expectation on the rhs approximated by a time average and that on the lhs by averaging over the ensemble and over time (e.g., Mitchell and Houtekamer 2000; Wang and Bishop 2003; Aksoy et al. 2009). Dee (1995) discusses in detail the use of Eq. (1) for covariance estimation.

Figure 2 depicts, for the first 10 days of June 2008, the square root of the rhs of Eq. (1) with $P$ estimated from the prior ensemble, which we term the total spread, and the rms innovations [i.e., the square root of the lhs of Eq. (1)]. With the NCEP observation errors, the rms innovations between the 3-h ensemble mean forecast and METAR observations of U10 were much smaller than the total spread (gray dashed line around 3.6 m s$^{-1}$). When we reduced the observation-error variance from $(3.5 \text{ m s}^{-1})^2$ to $(1.5 \text{ m s}^{-1})^2$, the forecast fit to the observations improved by $\sim 0.6 \text{ m s}^{-1}$ on average, making the innovation comparable to total spread. The $v$ wind at 10 m yielded very similar results when the same observation

![Figure 2](image-url)
errors were applied. For 2-m temperature, rms innovations improved more than 0.5 K with the observation-error variance reduced from \((2.5 \text{ K})^2\), as specified in the NCEP error table, to \((1.5 \text{ K})^2\) (not shown). Based on these experiments, those reduced observation-error variances were applied for all the subsequent experiments.

Moisture observations were assimilated as dewpoint temperature regardless of the observation type, following Fujita et al. (2007). As the MADIS provides moisture observations as dewpoint, we could avoid an additional conversion error in the observation operator and take advantage of the Gaussian characteristics of the observation and background errors for dewpoint.

The dewpoint observation-error variance was specified following Lin and Hubbard (2004). They estimated dewpoint errors based on the sensitivity of dewpoint to small perturbations in temperature and relative humidity, with the result that the error variance depends on the relative humidity and temperature and increases as the relative humidity decreases. There are three parameters to be specified: assumed uncertainties for temperature and relative humidity, and a minimum value allowed for the relative humidity, since the sensitivity of dewpoint to relative humidity becomes exponentially large as relative humidity approaches zero. Figure 3 shows histograms of the estimated dewpoint observation-error variance when the assumed relative humidity uncertainty and the minimum relative humidity are \(\pm 10\%\) and \(10\%\), respectively, or \(\pm 5\%\) and \(20\%\), while the assumed temperature uncertainty remains as \(\pm 1\text{ K}\). In the former case, most dewpoint observations are assigned error standard deviations between 2 and 3 K but the maximum error can be as large as 14 K when relative humidity becomes small (in the transparent bar graph). In the latter case, the observation-error standard deviations generally lie between 1.4 and 2.2 K and are no larger than 4 K. We summarize in Table 1 how these choices for the error variances affect the performance of the forecast over the 10-day test period. Averaged over roughly 1000 METAR stations in the CONUS domain, the larger error variances result in total spread that is substantially larger than the rms innovations for 2-m dewpoint. The smaller error variances reduce the rms innovations by 10% and give a total spread that matches the rms innovations.

Based on this result, we applied the new error specification regardless of vertical levels or instrument types in all experiments. This was done partly because of simplicity and partly because our main focus was on surface data assimilation, but Fig. 4 indicates that the mean bias and the rms errors of the 3-h forecast (i.e., prior) are greatly reduced in the entire atmosphere by reducing dewpoint observation errors at all levels. The improved moisture assimilation, however, does not noticeably affect the horizontal wind or temperature in the background forecast (not shown).

Surface observations are subject to considerable errors of representativeness, especially in regions of complex terrain. This means that the variance of observation error, which includes both the instrumental error and errors of representativeness, is uncertain for some surface observations. Moreover, innovation statistics for surface observations often exhibit significant bias, which may arise either from deficiencies in the forecast model or peculiarities of where the observations are located and that bias is not accounted for in the formulation of the assimilation system. We addressed these issues in only the simplest way, by assimilating just those observations for which the station elevation was within a certain distance from the model topography. Given that our

<table>
<thead>
<tr>
<th>Old</th>
<th>New</th>
</tr>
</thead>
<tbody>
<tr>
<td>rmse</td>
<td>2.2</td>
</tr>
<tr>
<td>Spread</td>
<td>3.6</td>
</tr>
</tbody>
</table>

Fig. 3. Histogram of the observation errors for METAR dewpoint temperature sampled from domain 1 at 0000 UTC 1 Jun 2008. The original error distribution estimated from Lin and Hubbard (2004) is shown as transparent bar graph while the newly adjusted error distribution in bar graph filled in black.
horizontal resolution (for either 45- or 15-km grid) does not capture the scales associated with narrow valleys, we tested two different threshold values of 100 and 300 m in rejecting surface observations. Compared to a more generous threshold of 300 m, the use of 100-m threshold discarded more surface observations (up to 14%), predominantly over the western United States; over the Great Plains, however, most observations were retained (Fig. 1). The verification of 3-h forecasts against surface METAR observations for the two different thresholds is summarized in Table 2. For the 10-day test period, rms innovations were not significantly reduced by rejecting more stations with the 100-m threshold, while the cold bias was decreased in 2-m temperature by about 20%.

As our observation-space diagnostics showed no consistent advantage for one threshold over the other, we chose the 300-m threshold in our assimilation to benefit from the dense observing network at the surface.

4. Results

Following the careful specification of observation errors and the filter design for mesoscale cycling, we ran experiments with and without surface observations for the month of June 2008. The experiments are named SFCDA and NO_SFCDA, respectively, and both use all other conventional observations discussed in section 3.

Figure 5 shows the diagnostics for the two runs in terms of rms innovations, ensemble spread, and model-minus-observation bias errors verified against independent mesonet observations in domain 1. With the assimilation of surface observations (SFCDA), both the prior and posterior ensemble mean fit the observations better throughout the period. A similar advantage of SFCDA is also clear in the METAR verification where both experiments produced better fits than in the mesonet verification by ~10% (not shown).

Among all surface fields, the largest improvement from the assimilation of surface observations is made in dewpoint temperature; the rms difference of the prior forecast from the observations decreases by 0.4 K when averaged over all the common stations between the two experiments for the entire cycling period (Fig. 5c). The rms innovations and the bias of the 3-h ensemble mean forecast, and ensemble spread, averaged over the cycles, are summarized in Table 3.

A large portion of rms innovations in the surface thermodynamic fields comes from bias. In comparison to mesonet observations, the 3-h forecast from NO_SFCDA presents cold and wet bias during the daytime while it is slightly warmer and much drier in stable boundary conditions. The assimilation of surface observations effectively decreases the magnitude and diurnal variability of the mean bias near the surface. The 3-h forecast from the SFCDA reduces the cold bias by up to 0.4 K in the

| Table 2 | The rms innovations and bias errors of the prior (i.e., 3-h forecast) verified against METAR observations computed over all the stations in domain 1 for the first 10 days of June 2008. The errors are compared between two experiments with different threshold values—dh100 for 100 m and dh300 for 300 m—as the maximum height difference between the stations and the model height to be assimilated. In the first column, “Nobs” stands for the total number of stations assimilated. |
|---------|-------------------------------------------------|-------------------------------------------------|-------------------------------------------------|-------------------------------------------------|-------------------------------------------------|-------------------------------------------------|
|         | U10 (m s⁻¹) dh100 dh300 | V10 (m s⁻¹) dh100 dh300 | T2 (K) dh100 dh300 | Td2 (K) dh100 dh300 |
| rmse    | 1.93 | 1.98 | 1.95 | 2.00 | 2.00 | 2.01 | 1.94 | 1.98 |
| Bias    | 0.48 | 0.54 | 0.38 | 0.30 | −0.26 | −0.32 | 0.15 | 0.14 |
| Nobs    | 936 | 1082 | 932 | 1079 | 961 | 1118 | 973 | 1129 |
afternoon and dry bias by up to 1.6 K during nighttime (Figs. 5b and 5c). These bias errors in the surface fields may have various causes including the representativeness error due to the smoother model terrain height than the actual orography, errors in the soil initialization, systematic deficiencies of the surface-layer scheme in representing the surface exchange coefficients (which are then used to compute surface fluxes), the LSM in representing the evolution of soil states, and the PBL scheme in parameterizing the subgrid-scale turbulent mixing. Despite all these different error sources, it is evident that the assimilation of surface observations effectively alleviates the surface thermodynamic bias errors (with the maximum time-mean decrease of ~0.5 K in 2-m dewpoint).

The spread of the 3-h ensemble forecasts is larger in NO_SFCDA than in SFCDA, but compared to the specified observation-error standard deviation (as described in the previous section), the spread difference between the two experiments is not considerable. That is, the observation uncertainty is dominant over the difference in the forecast uncertainties between the two runs.

Diagnostics in the nested domain illustrate similar advantages in SFCDA to the ones in domain 1 in all surface fields. Moisture assimilation makes the biggest improvement again—about 0.8 K in rms innovation and 0.7 K in bias errors on average—as shown in Fig. 6. Improvements from the assimilation of surface observations are generally larger in domain 2 partly because it mostly covers smooth terrain and partly because innovations are computed using the atmospheric states from the finest grid as discussed in section 2.

Another interesting aspect in the ensemble data assimilation system is to compare the number of observations actually used in each experiment. In general, our experiments use 80%–90% of surface observations depending on the observation type, both in the assimilation (e.g., for METAR) and in the verification (e.g., for mesonet observations). Without assimilation of METAR observations, the system rejects more mesonet observations (up to 10%) especially when the model has a strong bias with respect to the surface observations due to the failure of outlier check described in section 2b. The fact that SFCDA retains more observations is another sign that the SFCDA analyses and forecasts are improved relative to those from NO_SFCDA.

To examine the vertical impact of the assimilation of surface data, we plot vertical profiles of the rms forecast differences from radiosonde observations for both experiments in the nested domain in Fig. 7. With the assimilation of surface data, the 3-h ensemble mean forecast shows better fits to sounding observations as well as improved spread-error relationships for all four fields in the entire atmosphere. The superiority of SFCDA is also evident in domain 1, though it is slightly smaller (not shown). The rms innovations from radiosonde observations decrease by up to 0.4 m s\(^{-1}\) for horizontal winds at 850 mb and up to 0.4 K in temperature at 925 mb. Bootstrapping over 3000 resamples, these differences between the two experiments are statistically significant at 95% confidence intervals. The positive impact of the surface data assimilation is not limited to the surface or the boundary layer but improves the forecasts of the entire troposphere in all variables.
Now we return to the surface bias of the analyses and short-range forecasts. As shown in Fig. 5, the assimilation of surface observations reduces the forecast bias errors in surface thermodynamic fields more effectively than in surface winds, we therefore focus on temperature and dewpoint at 2 m.

Figure 8 illustrates the time-mean innovation at each METAR station in the SFCDA experiment. Surface temperature (Fig. 8a) and dewpoint (Fig. 8b) in the 3-h ensemble mean forecast are too warm and slightly dry in the central United States but are colder and wetter than observations elsewhere. To see how the bias varies through the diurnal cycle, we plot a time series of the bias averaged over all stations at different UTC cycles in Fig. 8c. It shows prominent warm and dry bias near the surface during nighttime (from 0300 to 1200 UTC) while the cold and wet bias is dominant during daytime (from 1500 to 0000 UTC). This implies that the amplitude of the diurnal cycle for surface temperature forecast is generally underestimated and the moisture diurnal cycle is out of phase between the forecast and the observations for this summer month.

These bias errors are consistent with those found in the previous studies. Shin and Hong (2011) compared five different PBL schemes available in WRF and found the warm bias in the stable boundary condition common in all the PBL parameterizations owing to the underestimation of the surface cooling rate. The cold and wet bias in the convective boundary layer was also reported in Hu et al. (2010) for three different PBL schemes in WRF including the YSU PBL during the summertime. In a similar cycling study using WRF–DART for mesoscale assimilation over the CONUS domain, Romine et al. (2013) also found colder and wetter surface bias with the Mellor–Yamada–Janjic (MYJ) scheme (Janjic 1994) than with YSU in their short-range forecasts. The reason for the under-prediction of surface temperature in the convective regime is, however, still not clear. As well as the model errors in the surface layer and the PBL schemes and the imperfection of the surface diagnostic algorithm, the initialization of ensemble soil states during the cycling can also be an important factor in characterizing the diurnal variation of thermodynamic bias errors near the surface (S.-Y. Hong and J. Dudhia 2013, personal communication). As described in section 2b, our soil states freely evolve throughout the month-long period. Further investigation is required for this significant issue on how to properly treat the soil states in the assimilation of surface observations in the cycling context.

In terms of surface winds, there was stronger southerly wind over Texas and stronger westerly wind over the Rockies compared to METAR observations (not shown). But as illustrated in Fig. 5 and Table 3, SFCDA had little effect on the domain-averaged surface wind bias in the month-long statistics.

To confirm that the assimilation of surface observations updated the thermodynamic fields in the right direction, Fig. 9 shows time-mean analysis increments (i.e., analysis minus background) in the SFCDA experiment. The horizontal distribution of the time-mean analysis increments (Fig. 9a) matches well with that of the mean bias errors (Fig. 8a) but with opposite sign, which demonstrates that SFCDA tends to reduce the bias errors in the background forecast. Separating the cycles according to the time of day shows that the analysis decreases the warm and dry bias during the nighttime (Fig. 9b) and reduces the cold and wet bias before the PBL collapses, especially in the western United States (Fig. 9c).

We also investigated the systematic impact of the surface changes on the boundary layer structure by computing the ensemble- and time-mean vertical profiles of potential temperature across the domain. At the

---

**TABLE 3. Mesonet verification of 3-h forecast, averaged over all common stations in domain 1 for the month of June 2008. SFC stands for SFCDA and NOSFC represents NO_SFCDA.**

<table>
<thead>
<tr>
<th></th>
<th>U10 (m s(^{-1}))</th>
<th>V10 (m s(^{-1}))</th>
<th>T2 (K)</th>
<th>Td2 (K)</th>
<th>Altimeter (hPa)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SFC</td>
<td>NOSFC</td>
<td>SFC</td>
<td>NOSFC</td>
<td>SFC</td>
</tr>
<tr>
<td>rmse</td>
<td>2.09</td>
<td>2.19</td>
<td>2.06</td>
<td>2.14</td>
<td>2.05</td>
</tr>
<tr>
<td>Bias</td>
<td>0.78</td>
<td>0.78</td>
<td>0.17</td>
<td>0.10</td>
<td>−0.06</td>
</tr>
<tr>
<td>Spread</td>
<td>0.56</td>
<td>0.75</td>
<td>0.57</td>
<td>0.78</td>
<td>0.48</td>
</tr>
</tbody>
</table>

---

**Fig. 6.** As in Fig. 5c, but for all the mesonet stations used in each experiment in domain 2 at 15-km grid resolution. Ensemble spread is not depicted.
location of the large positive analysis increment at 2100 UTC [marked as a star (★) in northwestern New Mexico in Fig. 9c], the reduction of the cold surface bias in the analysis enhances the development of PBL and makes the whole PBL structure warmer in the 3-h forecast during the daytime (Fig. 10a). The corresponding wind profile indicates that turbulent mixing was much stronger throughout the convective boundary layer in SFCDA (not shown). These changes result in raising the PBL height by about 300 m as shown by the thin horizontal lines in Fig. 10a. The positive (negative) mean analysis increment at the surface in SFCDA (Fig. 9c) resulted in raising (lowering) the PBL height in the following forecast compared to NO_SFCDA, as shown in Fig. 10b.

Another important question is how long the influence of surface data assimilation lasts in forecasts. Therefore, we extended forecasts from the ensemble mean analyses valid at 0000 and 1200 UTC every other day for 72 h and verified them against METAR surface observations over ~1200 stations. In Fig. 11, averaged over the 30 extended deterministic forecasts, the assimilation of
surface observations improved the forecast over the first 6 h in 10-m wind and 2-m temperature and for up to 72 h in 2-m dewpoint (although we did not show all 72-h forecast lead times to zoom in the differences for the first 6 h). As shown in Fig. 5 and Table 3, forecasts of surface dewpoint fit observations better than any other variables through the cycles thanks to the more sophisticated
observation-error specification at each station, which resulted in the biggest and the longest improvement in the extended forecast. These differences were statistically significant up to at least 24-h forecasts at the 95% confidence intervals in our bootstrapping. Separating the forecasts from 0000 and 1200 UTC initialization (not shown) revealed that bias in either subset was as large as \( \pm 1 \) K in both 2-m temperature and dewpoint, but owing to their opposite signs, the bias averaged over both initialization times was small, as shown in Figs. 11c and 11d. Although their bias errors varied strongly over diurnal cycle, the quality of forecasts from the 0000 and 1200 UTC analyses were similar in terms of both rms and bias errors (not shown).

We have also examined the quality of precipitation forecasts in the experiments using the fractions skill score (FSS; Roberts and Lean 2008; Schwartz et al. 2009). Unlike traditional point-by-point verification approaches, the FSS relaxes the requirement that forecast and observed events match at the grid scale for a forecast to be considered statistically perfect. Computing the FSS requires specifying a radius of influence \( r \) and precipitation exceedance threshold \( q \). The number of grid boxes where both the forecast and observed precipitation exceed \( q \) within the radius \( r \) is determined and divided by the total number of grid points available within the radius. In other words, the points that are observed and forecasted at the same time are transformed into fractional values for the particular combination of \( q \) and \( r \). The FSS varies from 0 to 1, with a perfect forecast being 1.

Figure 12 shows the FSS for the 3-h accumulated precipitation forecast verified against NCEP stage IV precipitation data at each forecast lead time. Because the rainfall verification was done in domain 2, stage IV fields (with a native resolution of 4.7 km) are interpolated to the 15-km grid before computing the FSS. The radius of influence was chosen as 105 km since it is comparable to the actual resolvable scale (roughly six grid intervals) of the 15-km grid. Overall, the scores were not very sensitive to the radius of influence or to the initialization time of ensemble forecast. Here, we show the FSS for precipitation thresholds of 1.0 and 10.0 mm over all forecasts from the 0000 UTC initialization. Forecasts from SFCDA analyses produced higher FSS than the forecasts from NO_SFCDA analyses for both light and heavy rain cases up to 12-h lead time. Although FSS from SFCDA for the 10-mm threshold is degraded after 12 h, the benefits at shorter lead times together with evidence from the case study in the next section demonstrate a positive impact of surface data assimilation on the forecast.

5. A case study—9 June 2008

Next we examine how the assimilation of surface observations influenced the boundary layer structure and the following forecast through a severe convective case associated with a strong surface front situated across the model domain for several days. For this case study, we present all the forecasts from the 15-km grid (e.g., in D2).
a. Case description and experiment design

An upper-level trough approached over the Rockies and the surface low was located in central Colorado at 1200 UTC 8 June 2008 developing a strong surface cold front over the central plains (not shown). At that time, there was a strong southwesterly low-level jet (LLJ) ahead of the front reporting the maximum speed of 63 kt (32.4 m s\(^{-1}\)) at Topeka, Kansas. At 1800 UTC 8 June 2008, the cold front extended from western Iowa to the Oklahoma–Texas Panhandle. A squall line was initiated along the surface front in northwestern Kansas (Fig. 13a).

The strong southerly LLJ transported moisture and contributed to the instability with the mixed-layer convective available potential energy (CAPE) on the order of 2000 J kg\(^{-1}\) ahead of the frontal boundary. Within a couple of hours, the squall line has elongated into northern Oklahoma along the front. By 0000 UTC 9 June 2008, the squall line was further intensified in the southern plains, producing the maximum radar reflectivity of 75 dBZ in western Texas (Fig. 13b). This also resulted in high winds, large hail, and heavy rainfall reaching a maximum of 66 mm (6 h)\(^{-1}\) at Gage, Oklahoma. As the boundary layer began to cool and

**Fig. 11.** A time series of rms innovations (solid lines) and bias ($f - o$) errors (dashed lines) computed for extended forecasts from the mean EnKF analyses with (i.e., SFCDA) and without assimilation of surface data (i.e., NO_SFCDA), verified against ~1200 METAR stations over 30 cycles (twice daily, every other day) for the month of June 2008 in terms of (a) $u$ wind and (b) $v$ wind at 10 m and (c) temperature and (d) dewpoint at 2 m.
decouple in the evening, the LLJ increased across the southern plains into central Oklahoma. With veering wind shear and strong instability ahead of the slowly propagating cold front, the squall line merged into a severe mesoscale convective system (MCS) that persisted for the next 24 h (not shown).

To investigate how the assimilation of surface observations affected the initiation of the MCS, we selected the analysis cycle at 1800 UTC 8 June 2008, which is about 3 h before convective initiation over northern Oklahoma.

Isolating the effects of surface observations is subtle, because they influence the analysis directly, at the time they are assimilated, and indirectly, by contributing to improvements in the prior forecasts as the assimilation system cycles. Thus, we performed two additional experiments. The first experiment, named SFC$^2$, assimilated surface observations as well as all other observations but used the prior ensemble forecasts from NO_SFCDA (which were unaffected by previous surface observations). The second experiment, named NOSFC$^1$, excluded surface observations from the assimilation but used the prior ensemble forecasts from SFCDA (which depended on surface observations before the 1800 UTC cycle). That is, the experiment name itself indicates whether or not the surface data was assimilated in the 1800 UTC cycle, and the superscript implies that the background was generated from the analysis with (+) and without (−) the assimilation of surface observations in the previous cycles. Here the idea is to examine how much information we can get from the surface data assimilation when the prior is bad near the surface in the first experiment, then to check the importance of surface observations at the analysis when the background forecast already contains all the surface information from the previous cycles in the second experiment.

b. Influence of surface observations

Figure 14 shows equivalent potential temperature ($\theta_e$) at 2 m and the CAPE from the ensemble mean analysis in the two experiments. In SFC$^2$, there was a large horizontal $\theta_e$ gradient associated with the strong surface front ranging from central Kansas to the Oklahoma–Texas Panhandle. The large CAPE ahead of the surface front was indicative of a favorable condition for convective initiation at 1800 UTC (Fig. 14a). In NOSFC$^1$, even with the good background with all the previous surface observation information and the assimilation of all other upper-level observations, the surface temperature and moisture gradients were weaker than in SFC$^2$, as was the convective instability in the warm sector (Fig. 14b).

Figure 15 illustrates the ensemble mean analysis increment (i.e., posterior–prior) at the lowest model level in SFC$^2$, where negative temperature increments were as large as 3 K and horizontal wind vector increments were bigger than 3 m s$^{-1}$. The structure of these increments enhanced the surface cold front and displaced it farther to the southeast. In NOSFC$^1$, there was no noticeable increment in all the fields at the lowest level (not shown), indicating that the low-level analysis increments in SFC$^2$ were produced almost exclusively by surface observations.

Figure 16 shows the cross section of analysis increments in temperature and vertical velocity along the
line AB in Fig. 15. In SFC', the biggest increments appeared behind the cold front and spanned ~8 model levels up to 2-km height. Crucially, the horizontal and vertical extent of the analysis increments in SFC' were not predefined like in classical nudging techniques, but were objectively determined by ensemble covariances that reflected the mesoscale structure and dynamics of the PBL as simulated in the model. In terms of vertical velocity, strong and narrow positive increments occurred in the entire atmosphere above the front with the maximum up to ~7.5 cm s\(^{-1}\), while smaller negative increments were present behind the front at lower levels, consistent with the displacement of the cold front (Fig. 16a). Without surface data assimilation, upper-level observations still produced positive \(w\) increments in the upper troposphere but they were somewhat broader and
shifted downstream (Fig. 16b). More importantly, NOSFC\textsuperscript{1} failed to provide the significant temperature and wind changes necessary for the development of the surface cold front in the right place at this analysis time.

These results prove that surface observations play a dominant role in making a considerable difference near the surface and contribute to the correct timing and location of the surface front. The horizontal and the vertical range within which surface data affects in EnKF vary depending on the model error structure in the simulated PBL and the localization radius we specified based on the ensemble size and the observation network in use.

To understand the relative impact of the background, we also present the same increments from the original SFCDA experiment that had continuous cycling with all available observations including surface data (Fig. 16c). Note this figure does not show the actual fields but their increments. As the prior was already good (with all the surface information digested in the previous times), the increments were slightly smaller. One noticeable thing is that the SFCDA analysis showed cold air slightly shifted toward B in the downstream region and the vertical motion associated with the surface front was right ahead of the negative temperature increment. This indicates that the front was located about 30 km ahead in SFCDA and moving faster than in SFC\textsuperscript{−} when the background was consistent with the surface information in the previous time. The impact of background was thus not as big as SFC\textsuperscript{−} in this particular case.

The better analysis leads to better precipitation forecasts at longer forecast lead times. Figure 17 depicts the ensemble-mean accumulated rainfall (mm) for 6 h in SFC\textsuperscript{−}, NOSFC\textsuperscript{1}, and NCEP stage IV data at 0000 UTC 9 June 2008. The mean of the 6-h ensemble forecast at the 15-km grid simulated the squall line a bit broader than NCEP stage IV data (which was processed at 4.7-km resolution), but SFC\textsuperscript{−} accurately predicted its location near the Oklahoma–Kansas border. NOSFC\textsuperscript{1} predicted heavy rainfall in southern Iowa, but missed the squall line in northwestern Oklahoma that developed into the MCS for the next 24 h in the central plains. The ensemble mean forecast is superimposed by the probability of the heavy rainfall $\geq 50$ mm (6 h\textsuperscript{−1}) in the ensemble forecast in both experiments contouring every 10% in Figs. 17a,b. They indicate more than 5 members among 50 predicted the accumulation larger than 50 mm in the contoured area. As shown in Fig. 17b, most...
members overpredicted heavy precipitation over Iowa and missed the rainband over Oklahoma and Kansas in NOSFC\textsuperscript{+}. The original SFCDA and NO_SFCDA experiments showed a similar comparison in terms of the precipitation forecast, but were not presented here.

**FIG. 16.** Cross section of analysis increments in temperature (K, shaded) and vertical velocity $w$ (contoured every 2.5 cm s\textsuperscript{-1}) in (a) SFC\textsuperscript{-}, (b) NOSFC\textsuperscript{+}, and (c) SFCDA experiments. Negative increments in temperature and $w$ are represented as shading and contours with dashed lines, respectively.

**FIG. 17.** Accumulated rainfall from the mean of 6-h ensemble forecasts valid at 0000 UTC 9 Jun 2008 in (a) SFC\textsuperscript{-} and (b) NOSFC\textsuperscript{+} experiments, compared to (c) NCEP stage IV data. Probability (%) of accumulated rainfall over 50 mm (6 h)\textsuperscript{-1} in the ensemble forecast is contoured from 10\% every 10\% interval in (a) and (b).
6. Summary and conclusions

Surface observations (e.g., 2-m thermodynamic fields and 10-m winds) resolve mesoscale variations both spatially and temporally. This paper has explored assimilation of these surface observations with an EnKF employing WRF implemented via DART. The assimilation system cycles for one summer month of June 2008, producing analyses every 3 h over the contiguous United States and including an embedded nested domain at 15-km resolution.

Relative to experiments using only other conventional observations, assimilation of surface observations improves both analyses and short-term forecasts of surface variables. Improvement of analyses, evaluated by comparing against independent (e.g., unassimilated) mesonet observations, is as much as 16% in the components of 10-m winds, 15% in 2-m temperature, and 28% in 2-m dewpoint. Assimilation of surface observations also improves the fit of short-range forecasts to surface winds and temperature for 6 h, and for substantially longer hours for dewpoint temperature. These results are all statistically significant at the 95% confidence intervals.

In contrast to previous results with the EnKF, assimilation of surface observations improves forecasts aloft as well, as seen by reduced differences from radiosondes through at least 500 hPa, with the biggest improvement in horizontal winds at 850 hPa. These improvements translate into better precipitation forecasts, as measured by the FSS, through 12-h lead times.

Bias accounts for a significant portion of the forecast error for surface variables. In 2-m temperature, the absolute value of the time-averaged 3-h forecast difference from observations is greater than 0.5 K at most stations. The bias has a pronounced diurnal variation, whose phase is consistent with an underestimation of the diurnal cycle on average, and also varies coherently across the CONUS. Both these characteristics point to deficiencies in the forecast model as the source of the bias. Assimilating surface observations reduces biases in surface variables and also leads, through the analysis-forecast cycle, to accompanying systematic changes in the structure and depth of the PBL.

The benefits of assimilating surface observations come in part from the ability of the EnKF to produce analysis increments that are consistent with the structure and dynamics of the boundary layer. In addition to the systematic effects of surface observations on the depth of the PBL, our case study of a squall line illustrated how the EnKF used information in the surface observations to modify the intensity of the cold boundary layer behind the surface front and to displace the frontal boundary, together with the associated wind direction, to be consistent with observations. Assimilation of surface observations also increased convective instability ahead of the front. These changes in the PBL and mesoscale flow near the surface led to improvements in the subsequent precipitation forecast.

To see benefits from surface observations, some care is required in assigning observation-error variances. Comparing rms innovations against the square root of the sum of the ensemble variance and observation-error variance showed that the variances assigned to surface observations in the GSI were significantly too large in our system. Based on the innovation statistics, we reduced observation-error standard deviations to 1.5 m s⁻¹ and 1.5 K for 10-m winds and 2-m temperature, respectively, and modified the observation-error variance for dewpoint as a function of relative humidity and temperature. The new observation-error statistics improved not only the reliability of the ensemble system but also the quality of the following forecast, with the greatest enhancements following from the formulation of the error variance for dewpoint observations.

There are three issues that offer clear potential for further improvements in assimilating surface observations, which we leave as topics for future work. First, representativeness errors are particularly acute for surface observations. We have accounted for these only crudely and indirectly, by rejecting observations at which the station elevation differs from the model orography by more than 300 m.

Second, coupling to the land surface strongly influences atmospheric variables within the PBL but we have not updated or adjusted the land surface state during assimilation. This is likely far from optimal, as much previous work, beginning with Mahfouf (1991), has emphasized the role of surface observations in estimation of the land surface state. An important research problem is the development of techniques for jointly updating the atmospheric and land surface states when assimilating surface observations.

Finally, the magnitude of forecast bias relative to surface observations indicates that, not surprisingly, the forecast model has important deficiencies and errors. Forecast model errors affect our results by degrading both the 3-h mean prediction of the observations and the forecast covariances that are used in the EnKF, yet the assimilation system implemented here only implicitly accounts for those errors, through an adaptive inflation of the forecast covariances. Direct improvements to the model are certainly helpful and have been the focus of several recent studies with WRF-DART (Cavallo et al. 2011; Torn and Davis 2012). In addition, we are actively investigating various model error schemes for the EnKF and will report results in a subsequent study.
Acknowledgments. We thank David Dowell, Glen Romine, Jimy Dudhia, and Jeff Anderson for helpful discussions on this work. Observations were provided by NOAA/MADIS. We appreciate the support from Tim Hoar and Nancy Collins on the data processing in the DART team at NCAR, Glen Romine on the image processing, Dave Gill for the WRF preprocessing, and Craig Schwartz on the precipitation verification. This study was partially funded by the U.S. Air Force Weather Agency.

REFERENCES


