SENSITIVITY OF SURFACE AIR TEMPERATURE ANALYSES TO BACKGROUND AND OBSERVATION ERRORS

Daniel P. Tyndall[1] and John D. Horel
Department of Atmospheric Sciences, University of Utah
Salt Lake City, UT

Manuel S. F. V. de Pondeca
Science Applications International Corporation, National Centers for Environmental Prediction
Washington, D.C.

Submitted to: *Weather and Forecasting*
May 19, 2009

[1] Corresponding author address:
University of Utah
Department of Atmospheric Sciences
135 South 1460 East, Rm. 819
Salt Lake City, UT 84112-0110
Email: d.tyndall@utah.edu

Abstract

A two-dimensional variational method is used to analyze 2-m air temperature over a limited domain (4° latitude × 4° longitude) in order to evaluate approaches to examine the sensitivity of the temperature analysis to the specification of observation and background errors. This local surface analysis (LSA) utilizes the 1-h forecast from the Rapid Update Cycle (RUC) downscaled to a 5-km resolution terrain for its background fields and observations obtained from the Meteorological Assimilation Data Ingest System.

The observation error variance as a function of broad network categories and error variance and covariance of the downscaled 1-h RUC background fields are estimated using a sample of over 7 million 2-m air temperature observations in the continental United States collected during the period 8 May – 7 June 2008. The ratio of observation to background error variance is found to be between 2 and 3. This ratio is likely even higher in mountainous regions where representativeness errors attributed to the observations are large.

The technique to evaluate the sensitivity of the 2-m air temperature to the ratio of the observation and background error variance and background error length scales is illustrated over the Shenandoah Valley of Virginia for a particularly challenging case (0900 UTC 22 October 2007) when large horizontal temperature gradients were present in the mountainous regions as well as over two entire days (20 and 27 May 2009). Sets of data denial experiments in which observations are randomly and uniquely removed from each analysis are generated and evaluated. This method demonstrates the effects of overfitting the analysis to the observations.

1. Introduction

High resolution mesoscale surface analyses are increasingly becoming necessary for a variety of meteorological applications, such as mesoscale modeling and forecasting, dispersion modeling of air pollutants and hazardous materials, aviation and surface transportation, fire management, as well as climate applications. An analysis system designed to meet these needs requires as many conventional surface observations as possible to capture local scale weather features (Horel and Colman 2005). To meet these goals, the National Centers for Environmental Prediction (NCEP) has developed the Real-Time Mesoscale Analysis (RTMA), a 5-km surface analysis that uses a two dimensional variational (2DVar) technique to assimilate synoptic and mesonet conventional surface observations, as well as satellite-based winds over the oceans (de Pondeca et al. 2007).

This research began in 2006 to help evaluate the RTMA, with particular attention placed on the characteristics of the analyses in regions of complex terrain (Tyndall 2008). During this research, it was determined that it would be beneficial to evaluate appropriate specifications of error covariances for the RTMA's downscaled Rapid Update Cycle (RUC) background field, as the operational RTMA at that time used subjectively modified error covariances determined from the North American Mesoscale Model (NAM). Variational assimilation systems are dependent upon the background error covariances and observations error variances, which control how observation information is spread across model gridpoints. This optimally combines the observation and background field data sets into a continuous analysis grid. Without proper specification of these error covariances, under- and overfitting problems can cause significant degradation in analysis quality on local and regional scales (Daley 1991, 1997) which have been observed in RTMA fields.

Because of the complexity and computational cost associated with running the RTMA, a local variational data assimilation system was developed to evaluate methods to determine the impact of error covariances. This local surface analysis (LSA) will be described in the next section along with the approach to estimate the error covariances. A data denial technique is presented that relies upon a space filling Hilbert curve (Sagan 1994) to maintain the spatial heterogeneity of the distribution of the observations within the analysis domain. The sensitivity of air temperature analyses to the specification of the error covariances is illustrated in Section 3 using this data denial method for a particular case, 0900 UTC 22 October 2007 over the Shenandoah Valley/Shenandoah National Park area of northern Virginia. This case and region were chosen because they lead to one of the toughest challenges for a surface air temperature analysis: large horizontal temperature gradients in regions of complex topography arising from a surface-based radiational inversion (Myrick et al. 2005). The results from this single case are corroborated using hourly analyses during two days with differing synoptic conditions (20 and 26 May 2009).

A discussion related to the methodology and results follows in Section 4. It should be noted that our study is not directed towards identifying specific error covariance parameters appropriate for the RTMA. Rather, this study helps to define an efficient data denial approach for developers of the RTMA or other analysis systems undergoing development. In addition, our study is intended to improve understanding of some of the limitations of analysis systems. For example, some of our results may appear counterintuitive in that the “best” analysis may not necessarily be one that constrains the analysis strongly by the observations, as is often done by other operational analyses (e.g., MatchObsAll, Foisy 2003, Soltow and Cook, 2008).

2. Method

a. Local surface analysis

The LSA using a 2DVar assimilation method was developed for this research to minimize the computational cost of running large numbers of sensitivity experiments using a full analysis system over the entire continental United States (i.e., the RTMA). Since some of the most complex aspects of any data assimilation system are associated with the preprocessing and quality control of the data, the LSA was designed to use the RTMA's terrain, derived from USGS elevation datasets with the help of the pre-processing programs from the Weather Forecast and Research (WRF) model.

The LSA computes its 2-m air temperature analysis for the limited domain by using the General Minimum Residual method (GMRES; Saad and Schultz 1986) to solve the basic 2DVar analysis equations:

$$\left(\mathbf{P_b}^{\mathrm{T}} + \mathbf{P_b}^{\mathrm{T}}\mathbf{H}^{\mathrm{T}}\mathbf{P_o}^{-1}\mathbf{H}\mathbf{P_b}\right)\mathbf{v} = \mathbf{P_b}^{\mathrm{T}}\mathbf{H}^{\mathrm{T}}\mathbf{P_o}^{-1}\left[\mathbf{y_o} - \mathbf{H}\left(\mathbf{x_b}\right)\right] \quad (1)$$

$$\mathbf{x_a} = \mathbf{x_b} + \mathbf{P_b}\mathbf{v} \quad (2)$$

In these two equations, $\mathbf{y_o}$ is observation dataset, $\mathbf{x_b}$ is the background field, $\mathbf{H}$ is the linear forward operator used to transform analysis gridpoints to the observation locations, and $\mathbf{P_b}$ and $\mathbf{P_o}$ are the background and observation error covariance respectively. The term $\mathbf{v}$ is solved iteratively using the GMRES method to yield the analysis, $\mathbf{x_a}$.

The LSA utilizes a 13-km RUC 1-h forecast downscaled to 5-km resolution as its background field. A more complete description of the background downscaling procedure can be found in Benjamin et al. (2007) and Jascourt (2007), but the downscaling processes for the temperature fields most critical for this study are:

1. The RUC temperature fields at all vertical levels are bilinearly interpolated horizontally from the 13-km resolution to the 5-km grid.
2. Temperature grids are vertically interpolated to the height of the analysis terrain, in a manner depending on one of two conditions:
    - When the RTMA terrain is lower than the RUC13 terrain, the RUC13's lapse rate from the lowest 25 mb is multiplied by the distance between the two elevations and is added to the 2-m temperature from the RUC13 to yield the new 2-m temperature. In cases in which the low level lapse rate shows an inversion, the RUC13's 2-m temperature is used as is.
    - When the RTMA terrain is higher than the RUC13 terrain, the downscaling uses the RUC13's temperature at a height of 2-m above the downscaled terrain for the new background temperature.

While these processes do help to add small scale temperature features to the lower resolution background field, they can create unphysical temperature features. For example, when strong surface-based inversions are present, the background field may be too warm in the valleys, as the RUC's terrain will generally be higher than that of the NDFD terrain.

The LSA can be run over any subdomain in the continental United States. The 4°×4° domain used in this study is depicted in Fig. 1, which includes the Shenandoah Valley and a swath of the Blue Ridge and Appalachian Mountains. As typical with any analysis system, the NDFD 5-km terrain depicted in Fig. 1 fails to capture many of the small-scale, often important, local terrain features (not shown).

Observations used by the LSA include synoptic and aviation observations, surface mesonet observations from a variety of networks, Coastal-Marine Automated Network (C-MAN)

bouy observations, as well as ship observations that are obtained from the Meteorological Assimilation Data Ingest System (MADIS; Miller et al. 2005). All observations used must fall within a ±12 min time window centered about the analysis time, except for observations from the Remote Automated Weather Stations (RAWS) mesonet, which must fall within a time window beginning 30 min before and ending 12 min after the analysis validation time. The latter broader time window for RAWS observations reflects the need to allow for the fixed hourly data collection times outside of the ±12 min time window for many of those stations that are usually located in critical data sparse regions (Olsen and Horel 2007). Because this observation dataset was designed for an operational surface analysis system, operational time constraints often prevent many networks with relatively slow transmission times from being used in the analysis. MADIS provides a quality control check for all observations (TSP 88-21-R2, 1994), which is used by the LSA for its observation quality control (for the case study described here, MADIS quality control eliminated 2% of the 11,745 observations available across CONUS for the analysis hour).

As shown in Fig. 2, the number of observations available for any particular analysis time is far fewer than the number of gridpoints for which the analysis is computed whether considered over limited domains or for the nation as a whole. For example, there are approximately 700 observations available for the region of interest here while the analysis is computed over 6,384 gridpoints. In addition, the unbalanced distribution of observations across the region with dense coverage in the Washington D.C. metropolitan area and limited coverage elsewhere is common elsewhere around the country.

As shown in Fig. 2, the observations in this study were subdivided into four categories: METAR, primarily aviation and synoptic reports at airport locations; PUBLIC, an aggregation of

networks including the Automated Weather System (AWS) and the Citizen Weather Observing Program (CWOP); Remote Automated Weather Stations (RAWS), located typically in remote locations for fire weather applications; and OTHER, which includes all remaining surface mesonet observations (such as the Oklahoma Mesonet, the MesoWest network, and various transportation agencies). These categories were chosen based on similar characteristics and siting recommendations between observations in each group; i.e., observations in the RAWS category are located generally in complex terrain while the PUBLIC observations tend to be sited in developed areas near homes or schools. Because of their ubiquity and timeliness, the PUBLIC observations are an increasingly critical resource for the RTMA (e.g., 56% of the observations used in the RTMA nationwide were from PUBLIC observations for the time sampled for Fig. 2).

b. Background and observation error covariance

In variational data assimilation, it is often assumed that the observation errors at one location are uncorrelated with those at another, and hence, $\mathbf{P_o}$ has only diagonal elements determined by the magnitude of the observation error variance. For the LSA, observation error variances can be applied separately for each observation category (i.e., PUBLIC or RAWS). Because it is not practical to specify $\mathbf{P_b}$ uniquely for every pair of gridpoint locations (because of the large array size), assumptions are used by the LSA that lead to $\mathbf{P_b}$ being a sparse matrix in which only the background error covariances between pairs of nearby gridpoints are assumed to be related to one another and the diagonal elements are determined by the magnitude of the background error variance. The LSA specifies the background error covariance structure in terms of the background error variance (σ_b^2) and exponential functions of horizontal (r_{ij}) and vertical (z_{ij}) distance:

$$\rho_{ij} = \sigma_b^{\,2} \exp\left(-\frac{r_{ij}^{\,2}}{R^2}\right)\exp\left(-\frac{z_{ij}}{Z^2}\right) \qquad (3)$$

where ρ_{ij} is the background error correlation between gridpoints i and j, and R and Z are horizontal and vertical scaling factors that specify the decorrelation length scale that are determined empirically.

The determination of these observation and background error variances as well as the decorrelation length scales follows from a statistical analysis that is described by Myrick and Horel (2006) and used previously by Lönnberg and Hollingsworth (1986) and Xu et al. (2001). Using that method, observations and corresponding nearest values of the background fields during the 30-day period from 8 May-7 June 2008 were used to assess characteristics of the observation and background errors for the continental United States as a whole.

The method described by Myrick and Horel (2006) was used to estimate the structure of the background error covariance through correlations between innovations at one station with those at all nearby stations. Figure 3 shows one example of these correlations, computed between a station in Winchester, VA (KOKV), and those within the Shenandoah subdomain computed from all available observations in the 30-day period. The correlations tend to drop off sharply with distance and then remain above 0.3 for roughly 75 km. The spatial pattern of the background error correlation in the vicinity of KOKV using horizontal and vertical decorrelation length scales of 40 km and 100 m respectively is indicated by the shading in the left panel of Fig. 3. Generally, the estimates of the observed correlations tend to be smaller nearby and larger over longer separation distances than those specified by these horizontal decorrelation length scales. They also do not show evidence of the strong vertical decorrelation implied by Eq. (3), although few observations are available at higher elevations to help define that structure.

The example presented in Fig. 3 is one of the more than 11,000 estimates of the background error covariance that can be computed from the month-long sample of observations and background fields across the CONUS domain. As discussed by Myrick and Horel (2006), the covariance between observation innovations at two points separated by distance r can be accumulated over all of the observation-gridpoint pairs for the monthly sample:

$$\mathrm{cov}(r) = \overline{(o_1 - b_1)(o_2 - b_2)} \tag{4}$$

where the variables o and b correspond to the observation and background value, respectively. Then, using the same assumptions as Myrick and Horel (2006),

$$\mathrm{cov}(r = 0) = \sigma_o^{\,2} + \sigma_b^{\,2} \tag{5}$$

and

$$\mathrm{cov}(r) = b_1' b_2' = \sigma_b^{\,2} \rho_{ij} \tag{6}$$

where ρ_{ij} is defined typically isotropically as the first exponential term on the right hand side of Eq. (3).

The covariance of observation innovation as a function of distance r during the month long period computed for every location in the continental United States is shown in Fig. 4 for all observation types as well as separated into the four primary network categories. Key statistics are also summarized in Table 1. The covariance drops slowly as a function of horizontal distance and does not asymptote to 0, which suggests that the downscaled RUC background fields contain errors that remain correlated over distances of hundreds of kilometers. The error correlation over extremely long distances is not due to a systematic bias (not shown). Although the general behavior of the error covariance as a function of radius is similar for the METAR, PUBLIC, and OTHER categories, the RAWS stations exhibit a roughly linear dependence with distance. This

result suggests that the characteristics of the background errors in regions of complex terrain differ from those in other regions of the country.

Fitting a fifth-order polynomial to the innovation covariance values for horizontal distance greater than 5 km and extrapolating the curve back to r = 0 km allows an estimation of σ_b determined for all network types to be 1.2°C (filled symbols in the left margin of Fig. 4). Estimates of the RUC background error variance computed separately as a function of network type are also summarized in Table 1. The dependence of the results on data density is suggested by the lower estimates of background error variance for the PUBLIC and OTHER categories relative to the METAR observations. The higher estimate of background error variance for RAWS observations reflects representativeness problems associated with these observations generally being located in regions of complex terrain. The observation error variance can be estimated from the difference between the innovation covariance value at distance zero (filled symbols in Fig. 4 and the $\sigma_b^2 + \sigma_o^2$ column in Table 1) and the estimates of σ_b^2 using Eq. (6). Thus, σ_o for all stations in the continental United States is roughly 2.5°C. As might be expected, σ_o for METAR stations is estimated to be lower (2.1°C) than that for other network types. The larger observation error (3.2°C) for RAWS stations is also expected, since the observation error arises from both instrumental and representativeness errors. Even an observation with minimal instrumental error may not be representative of the unknown true value on the scale of the 5×5 km^2 analysis grid.

Figure 4 and Table 1 provide support for using a ratio of σ_o^2: σ_b^2 of 2:1, as well as increasing R and Z to provide a slower decorrelation of background errors with distance. For example, the right panel of Fig. 3 shows the specification of the background error correlation when R and Z are doubled in Eq. (3), which tends to broaden the error correlation in a manner

more consistent with that estimated from the observations in this subdomain. Sensitivity of the LSA to the specification of the observation to background error variance and decorrelation length scales will be examined in Section 3.

c. Data denial methodology

Data denial experiments have been routinely used to quantitatively evaluate objective analyses (Zapotocny et al. 2000). The analyses computed with the restricted set of observations are usually compared to the withheld observations, the background fields, or the control analyses using all observations to define measures of accuracy and uncertainty of the analysis system (Seaman and Hutchinson 1985).

The estimates of analysis accuracy and uncertainty can be sensitive to the approach used to randomly remove the observations from the analysis. Simply removing randomly every tenth observation is not optimal unless the observations are uniformly distributed spatially and exhibit equal observation error. Since observation networks in the continental United States tend to be clustered near urban areas (Fig. 2), a withheld observation in an urban area may have less of an impact on an analysis than if that observation was located in an area of low observation density (Seaman and Hutchinson 1985; Myrick and Horel 2008). Previous researchers have avoided this problem by taking into account the spatial distribution of the observations (de Pondeca et al. 2006) or by removing observations only from more randomly distributed networks (Myrick and Horel 2008).

The data denial technique used in this study was designed following the cross validation tool developed for the RTMA, which attempts to minimize the impact of nonrandom observation networks by using the Hilbert curve (de Pondeca et al. 2006). The Hilbert curve is a space filling

curve that occupies its entire domain, maintains spatial uniformity, and never overlaps upon itself (Sagan 1994). As an illustration of the steps required in generating the Hilbert curve, consider Fig. 5. The first step is simply to define the sample of observations to be used, in this case, an artificial sample of 37 locations scattered across the continental United States (panel 1). Next, the entire domain is converted into a unit domain, and then subdivided into four quadrants, which in turn are each successively subdivided into four smaller quadrants, and the process is repeated *n*-times until there is a maximum of one observation in each of the smaller subsections of the domain (panel 2). As evident in the second panel, many subsections may not contain an observation. The Hilbert curve (dark grey segments) is drawn through each subsection in the entire domain. Subsections of the domain that do not contain an observation (light grey dots) are ignored (panel 3). Next, the order in which the observations are located along the Hilbert curve is used to determine whether or not they are withheld from a specific analysis (panel 4). Here, every fifth observation (black squares) (starting from the lower left hand corner of the domain and skipping empty subsections) is removed from the observation data set and considered to be the withheld sample (panel 5). This approach leads to five unique verification data sets that are spatially uniform in the context of the observation density.

While Fig. 5 illustrates how a Hilbert curve can be computed for a small data set, that approach is too computationally expensive to be used for the domain used in this study. The computationally efficient data denial algorithm used by the LSA was based off of the same algorithm used by the RTMA (de Pondeca et al. 2006), which computes a base-4 Hilbert coordinate based on each observation's latitude and longitude. Observations are then binned into withholding groups based on their sequential order along the Hilbert curve, ignoring vertices for which no observations are available.

The Hilbert curve data withholding methodology was applied to each observation network separately, instead of grouping all observations together and using a single Hilbert curve. This latter method would have removed many more PUBLIC observations than observations in other networks because the PUBLIC network represents almost 80% of the observations in the Shenandoah Valley domain. Therefore, the impact of the other networks would have been quite difficult to assess if a single Hilbert curve was used for all observations. It should be noted that, as a result of using four Hilbert curves, the spatial uniformity aspect of the Hilbert curve is occasionally compromised. Our use of this methodology for this particular case study removes approximately 6 METAR, 58 PUBLIC, 1 RAWS, and 7 OTHER observations in each of the 10 data subsets.

d. Error statistics

Root-mean-square error (RMSE) was computed at both withheld and at all observation locations. Following Myrick and Horel (2008), the RMSE is defined as:

$$\mathrm{RMSE} = \sqrt{\sum_{j=1}^{M}\sum_{i=1}^{N}\frac{\left(a_{ij}-o_{ij}\right)^2}{MN}} \qquad (7)$$

where o_{ij} are the withheld (all) observations, a_{ij} are the analysis values at the nearest grid point to the observations, N is the number of withheld (all) observations in each of the $M = 10$ data denial experiments. Since each observation can belong to only one member j, the RMSE at withheld (all) observations is calculated using each station once (ten times). This RMSE estimate of analysis accuracy ignores observation error, and hence, should be viewed as a relative, not absolute, measure of analysis accuracy (Myrick and Horel 2008). Since the analyses are independent of the withheld observations, the RMSE computed using the withheld observations

is a more representative measure of analysis quality overall as well as representative of the quality of the analysis in data void regions.

The root-mean-square sensitivity is also used in this study to estimate analysis accuracy. Following Zapotocny et al. (2000) and Myrick and Horel (2008), the root-mean-square sensitivity of the analyses is defined as:

$$\mathrm{S} = \sqrt{\sum_{j=1}^{M}\sum_{i=1}^{L}\frac{\left(d_{ij}-c_i\right)^2}{ML}} \qquad (8)$$

where c_i is the i^{th} analysis value from the control analysis that uses all observations, d_{ij} is the i^{th} analysis value for the j^{th} data withholding experiment, and L = 6384 is the total number of analysis gridpoints. As discussed by Zapotocny et al. (2000) and Myrick and Horel (2008), the sensitivity indicates the magnitude of analysis change resulting from withholding data; a small value of S implies that the analysis is largely unaffected by the removal of the observations.

3. **Results**

a. Case study: 0900 UTC 22 October 2007

The data denial methodology is illustrated using a synoptic situation characterized by a strong surface-based radiational inversion that is typically difficult to analyze objectively in mountainous regions, since the surface temperature gradient can be very large with strong cold pools located in valleys adjacent to warmer conditions on surrounding slopes (Myrick et al. 2005). For this particular case (0900 UTC 22 October 2007) centered on the Shenandoah Valley in northern Virginia, the 1200 UTC atmospheric sounding launched from Sterling, VA (KIAD) within the domain exhibited a 12.6°C temperature increase within the lowest 460 m (not shown). The 13-km RUC 1-h forecast and downscaling procedure used to transform the forecast to the 5-

km grid leads to a number of mesoscale southwest-northeast oriented bands across Virginia as shown in Fig. 6a: higher temperature on the east of the domain, lower temperature to the east of the Blue Ridge Mountains, higher temperature over the background's approximation of the Blue Ridge Mountains, and generally lower temperature to the west of that range. The observations, however, help to provide greater detail in many locations, e.g., lower temperatures along the valley floor of the Shenandoah Valley and higher temperatures on nearby slopes, which an experienced forecaster would know to be typical of the conditions in other regions of the subdomain as well.

The LSA control analysis shown in Fig. 6b uses all of the available observations, observation to background error variance ratio of 1:1, and horizontal and vertical decorrelation length scales of 40 km and 100 m respectively. These choices of error ratio and decorrelation length scales are comparable to those used by the RTMA at that time. For clarity, only the interior 2°×2° region demarcated in Fig. 6a is shown in Fig. 6b and the remaining panels of Fig. 6. The difference between the control analysis and the background is shown in Fig. 6c. From Figs. 6b and 6c, the impact of using the mesoscale observations in this instance are: lower temperatures immediately to the east of the Blue Ridge Mountains and in many mountain valleys; a tendency to increase the temperatures along many of the slopes where observations are available; and an increase in temperature in parts of the Washington D.C. metropolitan area. While the control analysis is generally closer to what an experienced analyst might expect in regions where observations are available, the control analysis remains close to the background in data voids. For example, the lack of observations and rapid vertical decorrelation of the background error constrains the control analysis along the spine of the Blue Ridge Mountains to remain close to the background field. This results in nonphysical features, such as much colder temperatures along the southern crest and higher temperatures along the crest to the north.

An analysis using all observations but increasing the observation to background error variance ratio to 2:1 and lengthening the horizontal and vertical decorrelation length scales to 80 km and 200 m respectively is shown in Fig. 6d. These adjustments follow from the results presented in Section 2b for the entire continental United States and are not necessarily a priori optimal choices for this particular region or case. As expected, less "trust" of the observations leads to smaller differences in Fig. 6e between the analysis and the background field than those evident in Fig. 6c. However, using broader background decorrelation length scales leads to greater lateral and vertical influence of differences between observations and the background. For example, temperatures are increased along the crest of the southern section of the Blue Ridge Mountains.

b. Data Denial

As an illustration of the impact of removing 10% of the observations as part of the data denial procedure, Fig. 6f shows the difference between the control analysis depicted in Fig. 6b and an LSA analysis in which 10% of the observations are withheld randomly. Blue (red) areas in Fig. 6f indicate where the control analysis is colder (warmer) than the withheld analysis due to the use of the withheld observation increments (green numbers). Adding large observation innovations where no other ones are available nearby leads to large differences in Fig. 6f, e.g., using the -3.5°C innovation in the coastal plain near the southern border influences the analysis over an area defined primarily by the horizontal decorrelation length scale while the influence of the 5.7°C innovation near the lower-left edge of the domain is constrained further by local vertical terrain gradients. Generally, the impact of withholding large innovations is small near

Washington, D.C., since the availability of so many other observations in that area diminishes the effect of omitting a few of them.

The methodology described in Section 2d was used to evaluate objectively sets of 10 LSA analyses in which 10% of the observations are uniquely and randomly withheld from each analysis. These sets of analyses use different combinations of observation to background error variance ratios and horizontal and vertical decorrelation length scales. Since this research evolved from examining RTMA analyses; the decorrelation length scales and error ratios used by the RTMA at the time of the case study were used as the base values.

Table 2 summarizes the results obtained from the data denial experiments. The RMSE between the background values and all 700 observations within the 4°×4° domain was found to be 2.15°C. Using the base values (Experiment 1), the RMSE evaluated using all observations (i.e., 90% of which are used in each analysis and 10% are not) is lowered to 1.62°C. However, this overstates the improvement of the analysis relative to the background, since the RMSE is reduced by only 0.22°C when evaluated using the observations withheld from each analysis (center column). The left column can be viewed as a measure of the quality of an analysis in data rich regions while the center column is a measure of the quality of an analysis in data voids. Hence, one desirable feature of analyses is to have comparable RMSEs in both data rich and data poor regions; low RMSE where observations are plentiful and high RMSE elsewhere is an indicator of overfitting to the observations. This is only applicable when a spatially uniform withholding methodology is used, such as the one described in this study.

The magnitudes of the RMSE values in Table 2 are less important than the differences from one experiment to another, since all of the magnitudes could be increased towards that found for the background by designing the experiments to use a higher percentage of withheld

observations. It is also not relevant for this study to estimate the statistical significance of the small differences between the values in Table 2, as our goal is to demonstrate an approach, rather than define particular parameter values. Given those caveats, it is not surprising that the RMSE is lowest when evaluated using all observations for those experiments (3 and 4) when the observations are "trusted" more, i.e., by either decreasing the horizontal and vertical background error decorrelation length scales (Experiment 3) or decreasing the observation to background error variance ratio (Experiment 4). However, Experiment 7 suggests that making both adjustments at the same time leads to overfitting of the observations since the RMSE using all observations increases.

The RMSE based on the independent withheld observations is found to be lowest by changing the base values in 2 different ways, either: (Experiment 3) keeping the observation to background error variance ratio unchanged but shortening the horizontal and vertical background error decorrelation length scales or (Experiment 6) increasing both the error variance ratio and the decorrelation length scales. Since Experiment 3 (and similarly Experiment 7) reflect greater reliance on the observations and the discrepancies between the RMSE values in the left and center columns are relatively large, adjusting the base values in this manner is likely to result in overfitting to the observations. In contrast, Experiment 6 places more confidence in the background, yet the RMSE at the withheld locations does not differ that much from that found at all locations.

Since the RMSE is computed at only 700 of the 6,384 gridpoints within the 4°×4° domain, the sensitivity shown in the right column of Table 2 provides an analysis quality metric evaluated at every gridpoint computed over each of the 10 data withholding analyses. The value of 0.26 (0.22) °C for Experiment 1 (6) serves as a baseline for comparison and represents the

accumulation of the squared differences shown in Fig. 6c (6e) plus those outside of that interior 2°×2° subdomain. The sensitivity to withholding observations is reduced for those experiments (5 & 6) where the observation to background error variance ratio is increased and enhanced where the error variance ratio is decreased (compare Experiments 1 and 4 or Experiments 3 and 7). Similarly, the sensitivity is reduced if the decorrelation length scales are decreased (compare Experiments 1 and 3 or Experiments 6 and 5). As a general rule, the changes in sensitivity are larger due to changes in error variance ratio rather than to those in decorrelation length scales.

c. Further Evaluation

To demonstrate the applicability of our approach and results more generally, control analyses within the Shenandoah Valley domain were computed for each hour during two twenty-four hour periods: 20 May 2009, a synoptically quiescent period during which a high pressure system dominated much of the domain accompanied by a nocturnal inversion that mixed out during the afternoon, and 26 May 2009, a synoptically active period during which surface air temperatures within the domain were strongly influenced by the intermittent progression and retreat of a stationary front accompanied by heavy precipitation and high winds. Hence, an additional 3,696 data denial experiments were completed leading to the results presented in Tables 3 and 4. The RMSE and sensitivity values in those tables are accumulated over each of the 24 hour periods.

The lower averaged RMSE value for the background fields during the synoptically active period (2.25^{o}C) relative to that for the quiescent period (2.41^{o}C) confirm the aforementioned tendencies for larger analysis errors in mountainous regions arising from large temperature gradients along slopes. Further, these additional data denial experiments illustrate the increased

sensitivity to withholding observations if the analysis is constrained too tightly by the observations. For example, when the decorrelation length scales and the observation to background error variance ratio are halved (i.e., Experiment 7 relative to Experiment 2), sensitivity increases by 50% and RMSE values increase for both synoptic cases. Examining each synoptic situation separately, RMSE values at withheld observation locations (middle column values) differ by only small amounts when averaged over one set of 240 data denial experiments compared to another. However, when those values are compared to the RMSE values computed using all observation locations, the same conclusion obtained from the single case in the previous subsection is found, i.e., that a better analysis can be obtained by weighting the observations less and using a broader background error decorrelation length scale (Experiment 6). Similarly, the lowest sensitivities are found for Experiment 6.

4. Discussion

This effort was designed to provide guidance on procedures to help specify the observation error variance and the background error covariance used in NCEP's operational RTMA or other similar analysis systems. As with most modeling systems, the preprocessing steps required for the RTMA are the most complex to reproduce externally. This research was able to take advantage of all of the preprocessing done for the RTMA by simply downloading the background fields and observation files required. Estimates of the ratio of the observational error variance to that of the background error variance as well as the horizontal distance over which background errors remain well correlated with one another were determined using an observational method (Myrick and Horel 2006) that relied on over 7 million 2-m temperature observations over the entire continental United States during a month long period. The

observation to background error variance ratio was estimated to be higher than that used by the RTMA at the beginning of this study and background errors were found to be correlated over much larger horizontal distances. Based on these results, as well as other research by NCEP staff, the RTMA now uses parameter values comparable to those suggested by this study.

These results are also of interest to evaluate objectively the downscaled RUC background fields as well as differences between mesonets. Based on these comparisons, the quality of the RUC 1-hr temperature forecasts is judged to be quite good when examined on the scale of the entire continental United States over an entire month. Ongoing, routine documentation of estimates of the downscaled RUC background bias and error variance for all analysed variables would be useful to RTMA analysis users to help assess the quality of the analyses.

Networks that tend to be located in regions of relatively high observation density (e.g., METAR and PUBLIC) tend to have lower observation errors than those located predominantly in low density areas (e.g., RAWS and OTHER). The latter results from the larger representativeness errors (discrepancies between the spatial and temporal scales of the observations relative to that of the analysis grid) found especially in mountainous areas. Estimating observation variance as a function of network type also suggests, that at least for temperature, the errors of the PUBLIC observations are roughly the same as METAR observations. Concerns regarding the quality of PUBLIC observations have been raised in the past, often due to the challenges of siting instrumentation in urban areas. Based on the results of this work, the quality of PUBLIC temperature observations passing the quality control procedures in place for the RTMA appears comparable to that obtained from the surface aviation network.

The relatively high total count of surface observations available in the continental United States now (~12000) is somewhat misleading as so many of the observations used by the RTMA

are located in urban areas where the data density is often greater than that necessary to adjust the background field adequately. Since the surface analysis is only as good as the observations used in it, greater attention needs to be placed on identifying and obtaining access to additional data, particularly in regions of otherwise low data density. Staff at NWS offices around the country have been very successful at identifying local data assets and should be encouraged to continue to assist in these efforts. A recent report by the National Academy of Science (National Research Council 2009) has a number of important recommendations to facilitate a national "network of networks".

This study developed a simplified local variational analysis system that was used to examine techniques to evaluate the sensitivity of analyses to observation and background errors. The Hilbert curve method used to randomly assign stations to independent samples was applied to construct data withholding experiments. This technique was applied to a case study, as well as 48 other hourly analyses to evaluate the methodology. In the case study, the method presented here was applied for a synoptic situation known to be difficult to analyze objectively but for which an experienced analyst would subjectively be able to handle relatively easily: a strong surface inversion with large horizontal gradients in temperature along mountain slopes. For this case, selecting a smaller (larger) vertical (horizontal) decorrelation length scale to specify the background error covariance might have been expected a priori to yield the best temperature analysis. These choices would have limited the impact of large background errors arising from the strong vertical temperature gradients while recognizing that the errors in the background field at one valley or ridge top location would apply to other ones some distance removed. This study showed that allowing a broader horizontal influence of observation innovations did improve the analysis, which is also consistent with the earlier estimation of appropriate background error

horizontal decorrelation length scale based on the month-long sample of background grids for the nation as a whole. Repeating 10 data denial experiments over two 24-hour periods and 7 different sets of parameters (a total of 3,696 analyses) confirms that the likelihood of overfitting increases when the analysis is constrained too tightly to the observations.

Although it is not our intention to estimate analysis parameters applicable for the entire continental United States for all seasons entirely on the basis of our results, the approach developed here shows promise as a method to evaluate such attempts for operational applications. However, our experience in this area suggests that it may not be practical to parameterize the background error covariance as a function of horizontal and vertical separation in such a way that the analysis will be optimal for all synoptic situations and regions of the country. One of the advantages of the RTMA is that the background error covariance can be specified generally in terms of one or more characteristics of the background in order to alleviate overfitting. One approach may be to use a measure of boundary layer stability from the background field, i.e., assuming that locations and synoptic situations with similar stability will have similar background errors. That hypothesis would need to be tested in a manner similar to that developed here through examination of background error statistics stratified by boundary-layer stability over a large sample of synoptic situations.

Acknowledgements

Support for this project provided by the NOAA/NWS CSTAR program grant NA07NWS4680003. We would like to thank the anonymous reviewers for their thoughtful comments on the earlier version of this manuscript.

References

Benjamin, S., J. M. Brown, G. Manikin, and G. Mann, 2007: The RTMA background -- Hourly downscaling of RUC data to 5-km detail. Preprints, *18th Conference on Numerical Weather Prediction*, Park City, UT, 4A.6.

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

——, 1997: Atmospheric Data Assimilation. *J. Meteorol. Soc. Jpn.*, **75**, 319-329.

de Pondeca, M. S. F. V., S.-Y. Park, J. Purser, and G. DiMego, 2006: Applications of Hilbert Curves to the Selection of Subsets of Spatially Inhomogeneous Observational Data for Cross-Validation and to the Construction of Super-Observations. Preprints, *2006 American Geophysical Union Fall Meeting*, San Francisco, American Geophysical Union, A31A-0868.

——, and Coauthors, 2007: The Status of the Real Time Mesoscale Analysis System at NCEP. Preprints, *18th Conference on Numerical Weather Prediction*, Park City, Utah, 4A.5.

Foisy, T., 2003: The MatchObsAll Analysis System. *Western Regional Technical Attachment. No. 03-02.* Available at http://www.wrh.noaa.gov/wrh/03TAs/0302/

Horel, J., and B. Colman, 2005: Real-Time and Retrospective Mesoscale Objective Analyses. *Bull. Amer. Meteor. Soc.*, **86**, 1477-1480.

Jascourt, S., cited 2008: Real-Time Mesoscale Analysis: What is the NCEP RTMA and how can it be used? [Available online at https://www.meted.ucar.edu/loginForm.php?urlPath=nwp/RTMA.]

Lönnberg, P., and A. Hollingsworth, 1986: The statistical structure of short-range forecast errors as determined from radiosonde data. Part II: The covariance of height and wind errors. *Tellus*, **38A**, 137-161.

Miller, P. A., M. F. Barth, L. A. Benjamin, R. S. Artz, and W. R. Pendergrass, 2005: The Meteorological Assimilation and Data Ingest System (MADIS): Providing value-added observations to the meteorological community. Preprints, *21st Conference on Weather Analysis and Forecasting*, Washington, D.C., American Meteorolgical Society, P1.95.

Myrick, D. T., J. D. Horel, and S. M. Lazarus, 2005: Local Adjustment of the Background Error Correlation for Surface Analyses over Complex Terrain. *Wea. Forecasting*, **20**, 149-160.

——, and ——, 2006: Verification of Surface Temperature Forecasts from the National Digital Forecast Database over the Western United States. *Wea. Forecasting*, **21**, 869-892.

——, and ——, 2008: Sensitivity of Surface Analyses over the Western United States to RAWS Observations. *Wea. Forecasting*, **23**, 145-158.

National Research Council, 2009: *Observing Weather and Climate from the Ground Up: A Nationwide Network of Networks*. National Academy Press, 234 pp.

Olsen, B., and J. Horel, 2007: Sensitivity of Surface Analyses to Temporal Observational Constraints. Preprints, *18th Conference on Numerical Weather Prediction*, Park City, UT, P1.37.

Saad, Y., and M. H. Schultz, 1986: GMRES: A generalized minimal residual algorithm for solving nonsymmetric linear systems. *SIAM J. Sci. Stat. Comput.*, **7**, 856-869.

Sagan, H., 1994: *Space-filling curves*. Springer-Verlag, 193 pp.

Soltow, M., and K. Cook, 2008: Evaluation of the Real Time Mesoscale Analysis (RTMA) and the MatchObsAll (MoA) analysis in complex terrain. *Western Regional Technical*

Attachment. No. 08-01. 12 pp. Available from http://www.nwsla.noaa.gov/wrh/08TAs/ta0801.pdf

Seaman, R. S., and M. F. Hutchinson, 1985: Comparative real data test of some objective analysis methods by withholding observations. *Aust. Meteorol. Mag.*, **33**, 37-46.

Technique Specification Package 88-21-R2 For AWIPS-90 RFP Appendix G Requirements Numbers: Quality Control Incoming Data, 1994. AWIPS Document Number TSP-032-1992R2, NOAA, National Weather Service, Office of Systems Development.

Tyndall, D. P., 2008: Sensitivity of Surface Temperature Analyses to Specification of Background and Observation Errror Covariances, Dept. of Meteorology, University of Utah, 98 pp.

Xu, Q. L., L. Wei, A. V. Tuyl, and E. H. Barker, 2001: Estimation of three-dimensional error covariances. Part I: Analysis of height innovation vectors. *Mon. Wea. Rev.*, **129**, 2126-2135.

Zapotocny, T. H., and Coauthors, 2000: A Case Study of the Sensitivity of the Eta Data Assimilation System. *Wea. Forecasting*, **15**, 603-621.

List of Figure Captions

Figure 1. The Shenandoah Valley subdomain used in this study, with elevation shaded in meters. The inner box is a 2°×2° area, which will be focused on later in this paper. The sounding depicted in Fig. 2 was launched from the sounding site KIAD, labeled with an × in this figure.

Figure 2. Observation density across domain, grouped into four main categories. The first number in the upper right hand corner of each plot indicates the number of observations of the particular category located within the 2°×2° inner domain, the second number indicates the total number of observations within the entire 4°×4° domain, and the third number indicates the number of observations available across the entire continental United States for this particular analysis time.

Figure 3. Correlations (numbers) between temperature innovations at temperature observation Winchester, VA (KOKV) and all other locations within the 2°×2° Shenandoah subdomain computed over the 8 May – 7 June 2008 period. Shading indicates the shape of the background error correlation calculated by Eq. (3) with two different decorrelation length scales. The plot on the left uses horizontal and vertical decorrelation length scales of 40 km and 100 m; the plot on the right uses decorrelation length scales of 80 km and 200 m. Range rings are contoured in 25 km intervals and elevation is contoured in meters.

Figure 4. Binned innovation covariance (symbols) computed for the downscaled RUC background fields for the period 8 May 2008 to 7 June 2008 as a function of network type. Curve fits to the covariance are shown as a function of network type. The filled in symbols at $r = 0$ km

indicate extrapolated estimates of background error variance as a function of network type. The open symbols at $r = 0$ km denote estimates of the sum of the observation and background error variance computed as a function of network type. Background error covariance specified by Eq. (3) as a function of horizontal distance is also shown assuming lengths scales of 40 and 80 km (dotted lines).

Figure 5. Example of the Hilbert curve binning observations into data groups by withholding every 5th observation for a set of arbitrary observations across the continental United States.

Figure 6. Temperature analyses, increments, and background field used in this research. **a.** Temperature (°C) from the 1-hr RUC forecast downscaled to LSA resolution used as a background for the LSA over the entire 4°×4° domain. **b.** LSA temperature analysis (°C, shaded) over 2°×2° subdomain using horizontal and vertical decorrelation length scales of 40 km and 100 m and an observation to background error variance ratio of 1:1. Dots denote observation temperatures, which are colored according to the analysis temperature scale. **c.** Innovations (shaded, °C) of LSA temperature analysis in Fig. 3c. Dots indicate observation innovations, which are consistent with the gridded innovation scale. **d.** Same as Fig. 3b, but using horizontal and vertical decorrelation length scales of 80 km and 200 m and an observation to background error variance ratio of 2:1. **e.** Same as Fig. 3c, but depicting innovations of temperature analysis in Fig. 3d. **f.** Difference (°C) between control LSA temperature in which all observations are used and LSA temperature analysis in which only 90% of the observations are used for one withholding group over the 2°×2° subdomain. Purple dots indicate observations common to both analyses, while green numbers indicate the 10% of observation innovations withheld.

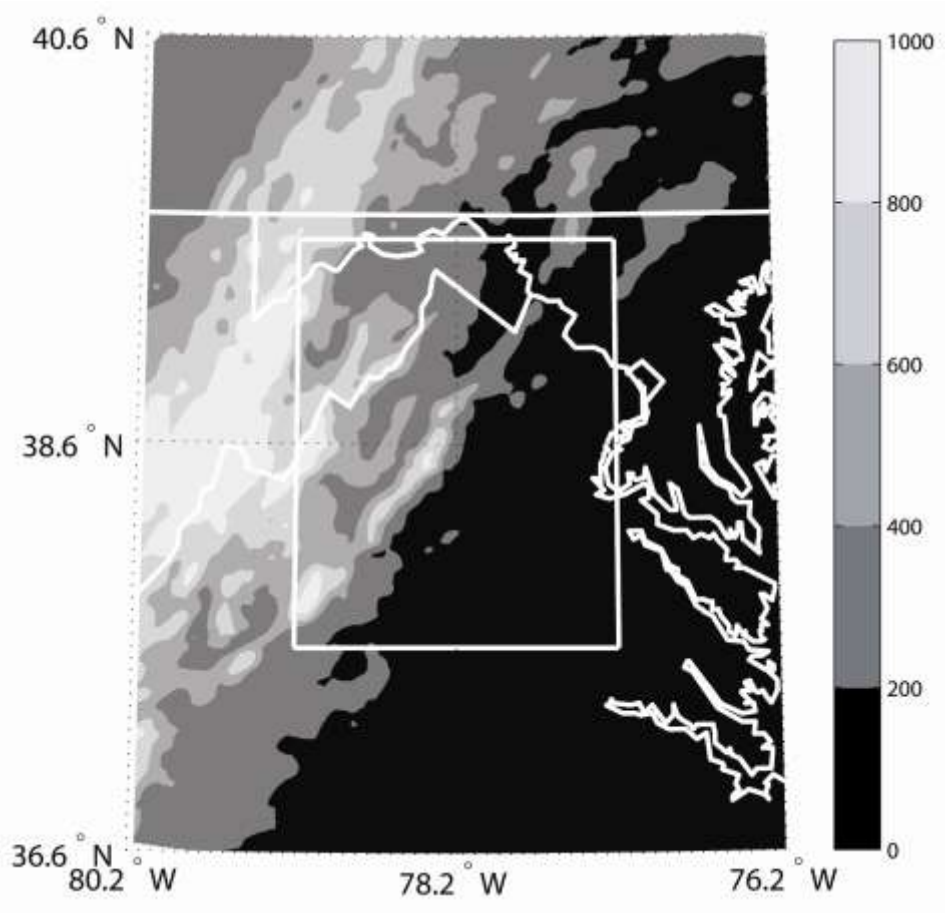

Figure 1. The Shenandoah Valley subdomain used in this study, with elevation shaded in meters. The inner box is a 2°×2° area, which will be focused on later in this paper. The sounding depicted in Fig. 2 was launched from the sounding site KIAD, labeled with an × in this figure.

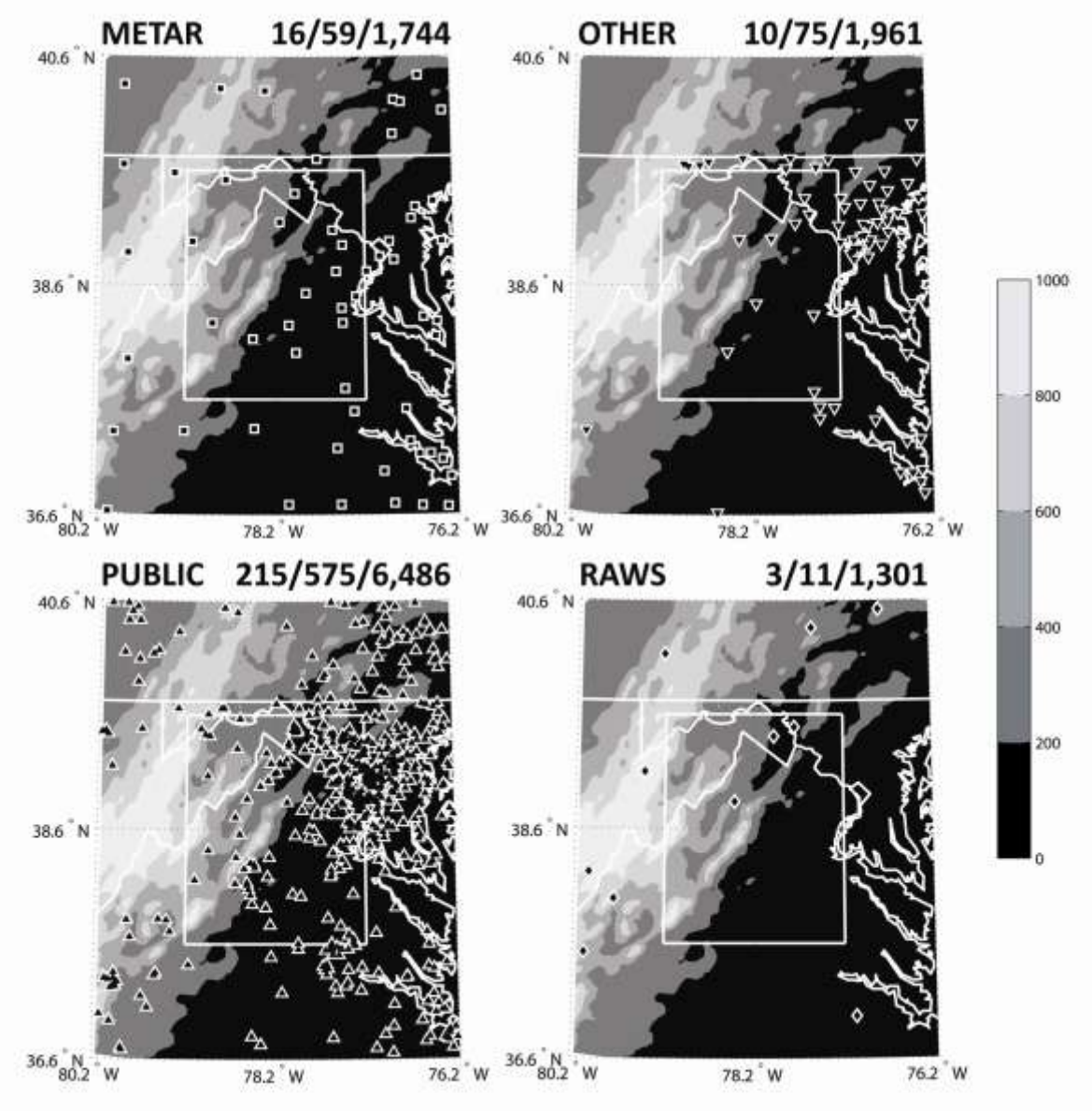

Figure 2. Observation density across domain, grouped into four main categories. The first number in the upper right hand corner of each plot indicates the number of observations of the particular category located within the 2°×2° inner domain, the second number indicates the total number of observations within the entire 4°×4° domain, and the third number indicates the number of observations available across the entire continental United States for this particular analysis time.

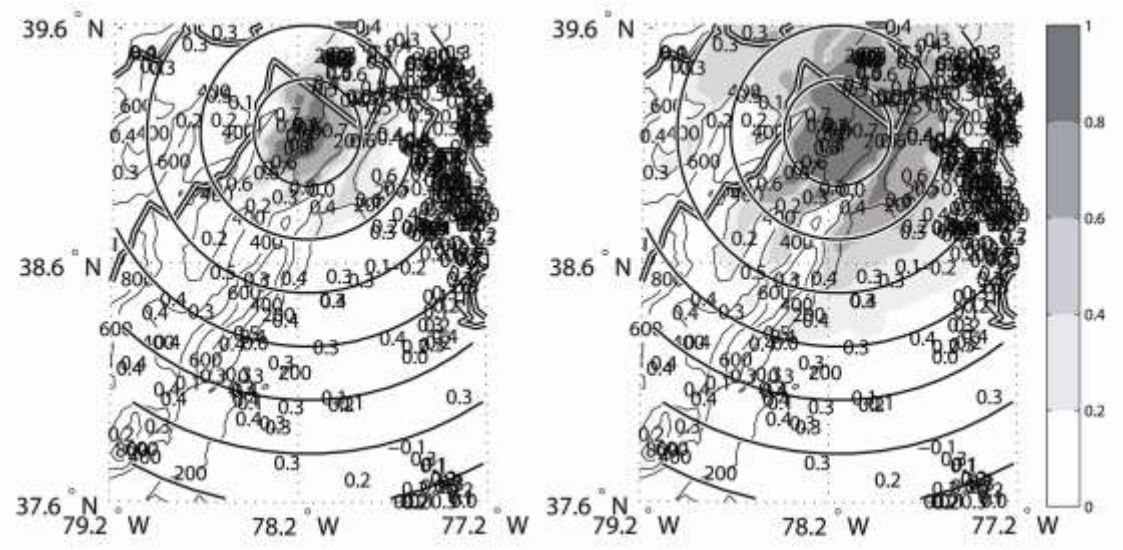

Figure 3. Correlations (numbers) between temperature innovations at temperature observation Winchester, VA (KOKV) and all other locations within the 2°×2° Shenandoah subdomain computed over the 8 May – 7 June 2008 period. Shading indicates the shape of the background error correlation calculated by Eq. (3) with two different decorrelation length scales. The plot on the left uses horizontal and vertical decorrelation length scales of 40 km and 100 m; the plot on the right uses decorrelation length scales of 80 km and 200 m. Range rings are contoured in 25 km intervals and elevation is contoured in meters.

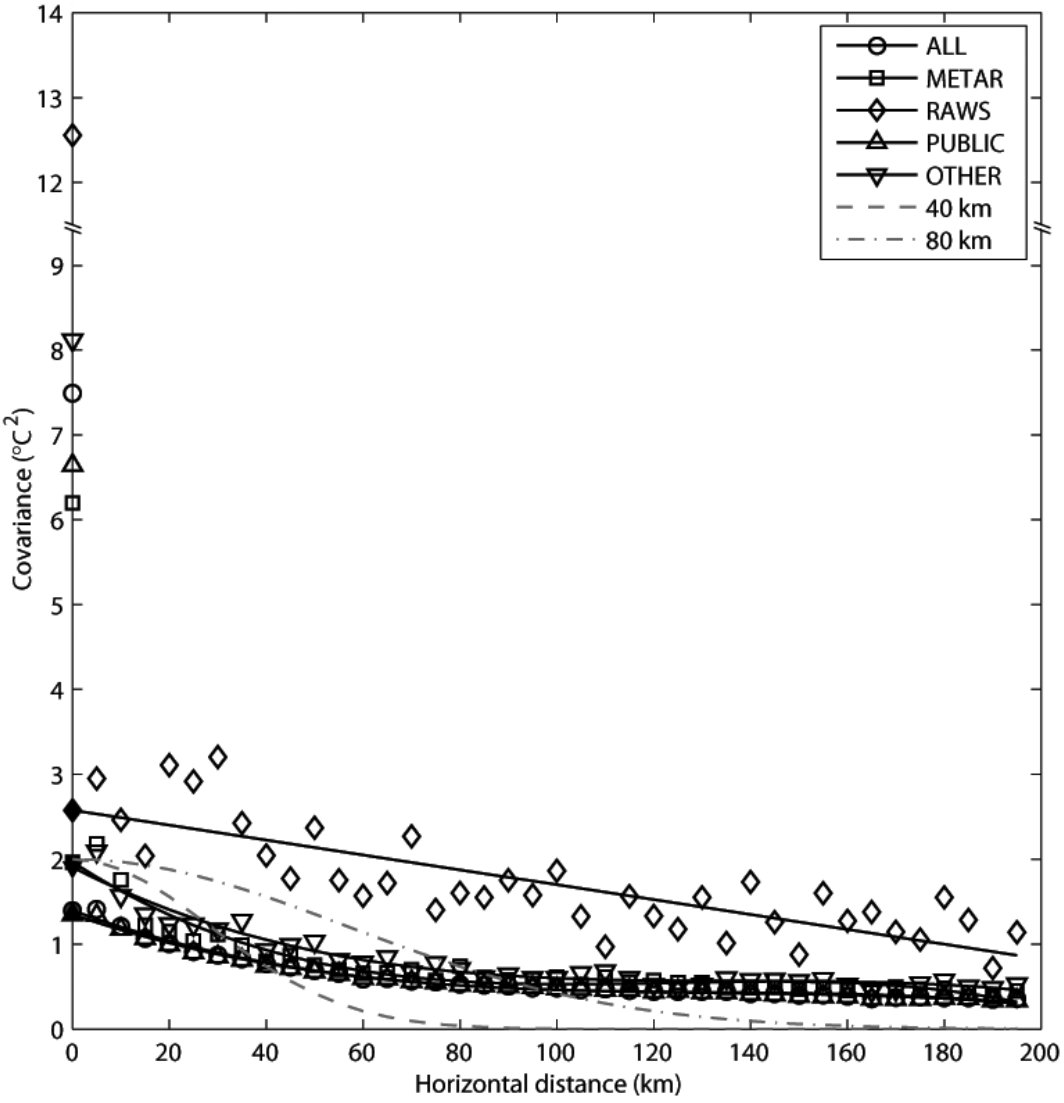

Figure 4. Binned innovation covariance (symbols) computed for the downscaled RUC background fields for the period 8 May 2008 to 7 June 2008 as a function of network type. Curve fits to the covariance are shown as a function of network type. The filled in symbols at $r = 0$ km indicate extrapolated estimates of background error variance as a function of network type. The open symbols at $r = 0$ km denote estimates of the sum of the observation and background error variance computed as a function of network type. Background error covariance specified by Eq.

(3) as a function of horizontal distance is also shown assuming lengths scales of 40 and 80 km (dotted lines).

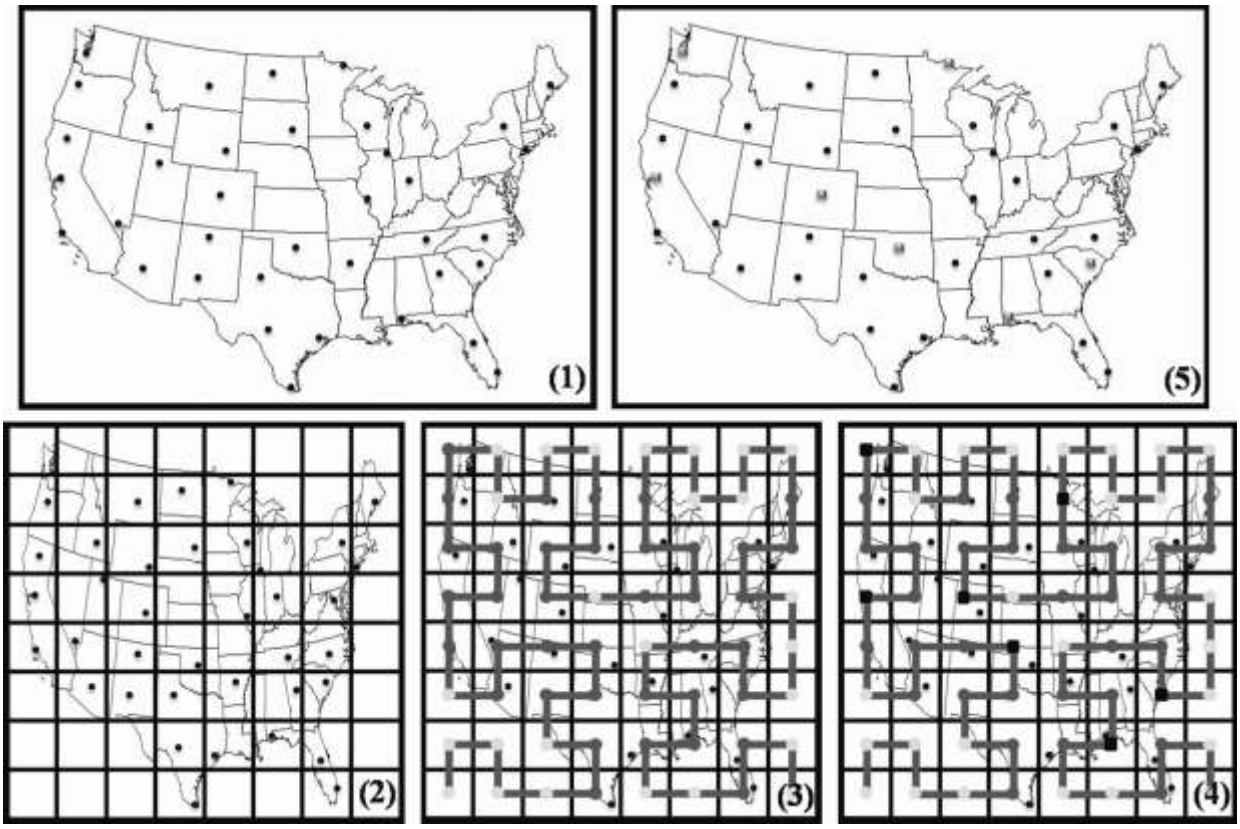

Figure 5. Example of the Hilbert curve binning observations into data groups by withholding every 5th observation for a set of arbitrary observations across the continental United States.

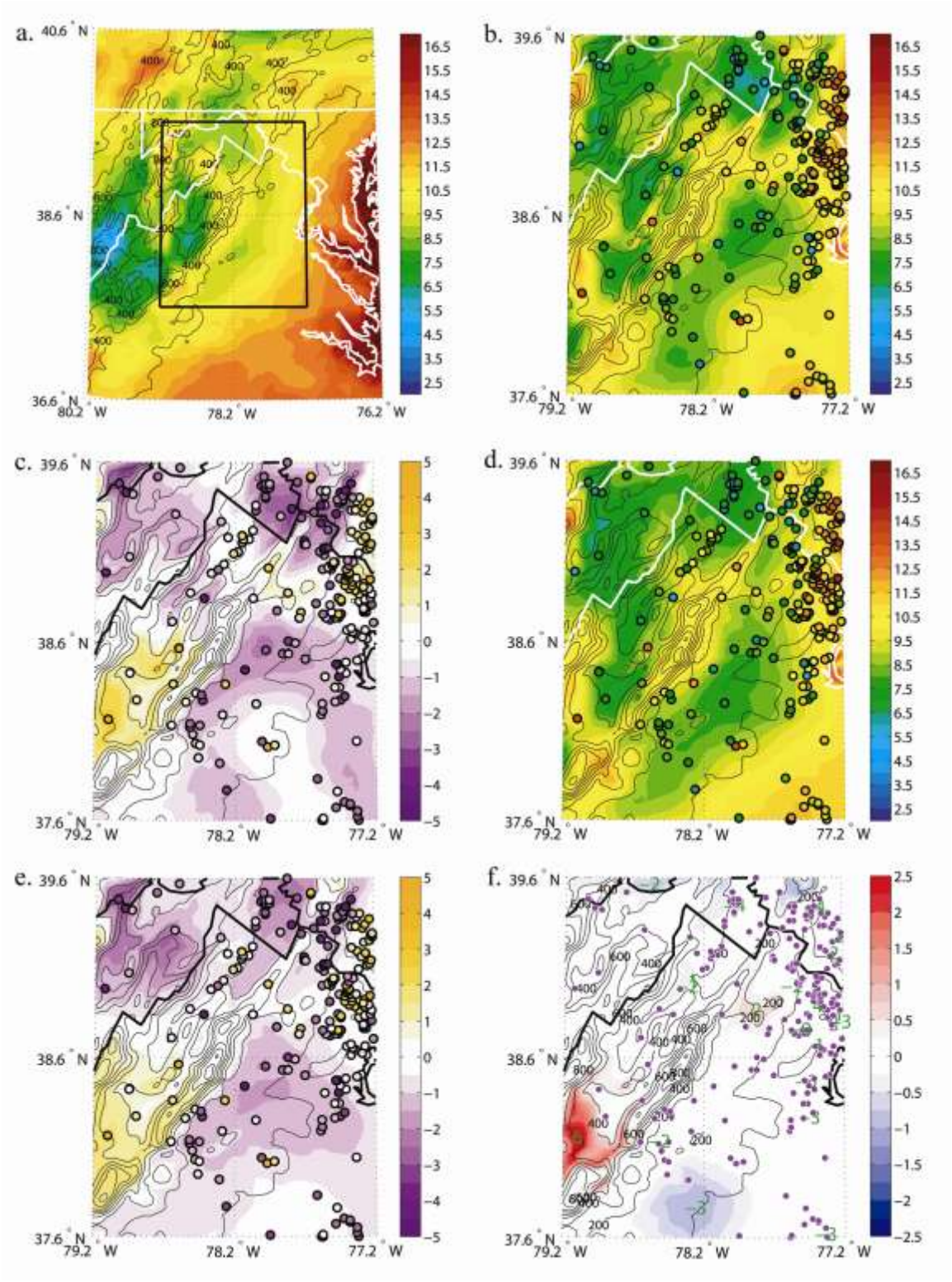

Figure 6. Temperature analyses, increments, and background field used in this research. **a.** Temperature (°C) from the 1-hr RUC forecast downscaled to LSA resolution used as a

background for the LSA over the entire 4°×4° domain. **b.** LSA temperature analysis (°C, shaded) over 2°×2° subdomain using horizontal and vertical decorrelation length scales of 40 km and 100 m and an observation to background error variance ratio of 1:1. Dots denote observation temperatures, which are colored according to the analysis temperature scale. **c.** Innovations (shaded, °C) of LSA temperature analysis in Fig. 3c. Dots indicate observation innovations, which are consistent with the gridded innovation scale. **d.** Same as Fig. 3b, but using horizontal and vertical decorrelation length scales of 80 km and 200 m and an observation to background error variance ratio of 2:1. **e.** Same as Fig. 3c, but depicting innovations of temperature analysis in Fig. 3d. **f.** Difference (°C) between control LSA temperature in which all observations are used and LSA temperature analysis in which only 90% of the observations are used for one withholding group over the 2°×2° subdomain. Purple dots indicate observations common to both analyses, while green numbers indicate the 10% of observation innovations withheld.

Table 1

Observation and Background Error Variances

Network	σ_b^2 **(°C^2)**	$\sigma_b^2 + \sigma_o^2$ **(°C^2)**	σ_o^2 **(°C^2)**	**Avg. num./hr**
ALL	1.4	7.5	6.1	11,464
METAR	2.0	6.2	4.2	1,744
PUBLIC	1.4	6.6	5.3	6,486
RAWS	2.6	12.6	10.0	1,301
OTHER	1.9	8.1	6.2	1,961

Table 2

RMSE and Sensitivity Over Shenandoah Valley Subdomain, 0900 UTC 22 October 2007

#	Experiment	RMSE Using All Observations (°C)	RMSE Using Withheld Observations (°C)	Sensitivity (°C)
B	Background	2.15	2.15	-
1	R = 40 km, Z = 100 m, $\sigma_o^2/\sigma_b^2 = 1$	1.62	1.93	0.26
2	R = 80 km, Z = 200 m, $\sigma_o^2/\sigma_b^2 = 1$	1.80	1.98	0.29
3	R = 20 km, Z = 50 m, $\sigma_o^2/\sigma_b^2 = 1$	1.41	1.89	0.20
4	R = 40 km, Z = 100 m, $\sigma_o^2/\sigma_b^2 = 0.5$	1.54	1.94	0.34
5	R = 40 km, Z = 100 m, $\sigma_o^2/\sigma_b^2 = 2$	1.70	1.93	0.19
6	R = 80 km, Z = 200 m, $\sigma_o^2/\sigma_b^2 = 2$	1.83	1.89	0.22
7	R = 20 km, Z = 50 m, $\sigma_o^2/\sigma_b^2 = 0.5$	1.67	1.90	0.26

Table 3

Accumulated RMSE and Sensitivity Over Shenandoah Valley Subdomain: 20 May 2009

#	Experiment	RMSE Using All Observations (°C)	RMSE Using Withheld Observations (°C)	Sensitivity (°C)
B	Background	2.41	2.41	-
1	R = 40 km, Z = 100 m, $\sigma_o^2/\sigma_b^2 = 1$	1.99	2.23	0.24
2	R = 80 km, Z = 200 m, $\sigma_o^2/\sigma_b^2 = 1$	2.12	2.22	0.19
3	R = 20 km, Z = 50 m, $\sigma_o^2/\sigma_b^2 = 1$	1.81	2.25	0.23
4	R = 40 km, Z = 100 m, $\sigma_o^2/\sigma_b^2 = 0.5$	1.94	2.25	0.30
5	R = 40 km, Z = 100 m, $\sigma_o^2/\sigma_b^2 = 2$	2.04	2.22	0.18
6	R = 80 km, Z = 200 m, $\sigma_o^2/\sigma_b^2 = 2$	2.14	2.22	0.16
7	R = 20 km, Z = 50 m, $\sigma_o^2/\sigma_b^2 = 0.5$	1.72	2.28	0.31

Table 4

Accumulated RMSE and Sensitivity Over Shenandoah Valley Subdomain: 26 May 2009

#	Experiment	RMSE Using All Observations (°C)	RMSE Using Withheld Observations (°C)	Sensitivity (°C)
B	Background	2.25	2.25	-
1	R = 40 km, Z = 100 m, $\sigma_o^2/\sigma_b^2 = 1$	1.81	2.01	0.21
2	R = 80 km, Z = 200 m, $\sigma_o^2/\sigma_b^2 = 1$	1.92	2.01	0.17
3	R = 20 km, Z = 50 m, $\sigma_o^2/\sigma_b^2 = 1$	1.65	2.03	0.21
4	R = 40 km, Z = 100 m, $\sigma_o^2/\sigma_b^2 = 0.5$	1.77	2.03	0.27
5	R = 40 km, Z = 100 m, $\sigma_o^2/\sigma_b^2 = 2$	1.85	2.00	0.16
6	R = 80 km, Z = 200 m, $\sigma_o^2/\sigma_b^2 = 2$	1.94	2.01	0.14
7	R = 20 km, Z = 50 m, $\sigma_o^2/\sigma_b^2 = 0.5$	1.58	2.06	0.27