A Comparison of Methods for Filling Gaps in Hourly Near-Surface Air Temperature Data

BRIAN HENN, MARK S. RALEIGH, ALEX FISHER,* AND JESSICA D. LUNDQUIST

Department of Civil and Environmental Engineering, University of Washington, Seattle, Washington

(Manuscript received 20 February 2012, in final form 7 September 2012)

ABSTRACT

Near-surface air temperature observations often have periods of missing data, and many applications using these datasets require filling in all missing periods. Multiple methods are available to fill missing data, but the comparative accuracy of these approaches has not been assessed. In this comparative study, five techniques were used to fill in missing temperature data: spatiotemporal correlations in the form of empirical orthogonal functions (EOFs), time series diurnal interpolation, and three variations of lapse rate–based filling. The method validation used sets of hourly surface temperature observations in complex terrain from five regions. The most accurate method for filling missing data depended on the number of available stations and the number of hours of missing data. Spatiotemporal correlations using EOF reconstruction were most accurate provided that at least 16 stations were available. Temporal interpolation was the most accurate method when only one or two stations were available or for 1-h gaps. Lapse rate–based filling was most accurate for intermediate numbers of stations. The accuracy of the lapse rate and EOF methods was found to be sensitive to the vertical separation of stations and the degree of correlation between them, which also explained some of the regional differences in performance. Horizontal distance was less significantly correlated with method performance. From these findings, guidelines are presented for choosing a filling method based on the duration of the missing data and the number of stations.

1. Introduction

Missing data are a common problem in meteorological and hydrologic observational datasets. Instruments may break, or data transmission may be interrupted. A meteorological phenomenon of interest (i.e., precipitation, snow, or ice) may even cause the temporary failure to record observations. Erroneous or unreasonable values may also be recorded and must be removed during data processing.

Thus, quality-controlled datasets often feature periods of missing data, a fact that may be problematic for reasons beyond the simple reduction in the number of data points. When calculating longer-term means using observed data, the absence of particular station data at a given time can introduce bias. Rolland (2003) found that lapse rate calculations relying on incomplete annual records produced unwanted variance because of the lapse rate’s seasonal variability. In the development of the Parameter-Elevation Regressions on Independent Slopes Model (PRISM) climatology product, Daly et al. (2008) required that at least 85% of the days in a month be present in a station’s record for that location to be included in the monthly analysis. On longer time scales, the installation and deactivation of temperature and precipitation observing sites require that measures be taken to avoid spurious trends (Maurer et al. 2002; Hamlet and Lettenmaier 2005). Detection of temporal trends in aggregate data requires that all stations have complete records or that methods be used to account for the changing set of stations over time.

Missing data are also unacceptable if the observations are to be used as the meteorological forcing data of a numerical model. For example, distributed hydrological models require that inputs be both temporally and spatially complete (Wigmosta et al. 1994; Daly et al. 2000), which may necessitate both filling of missing data

* Current affiliation: Horn Point Laboratory, University of Maryland Center for Environmental Science, Cambridge, Maryland.

Corresponding author address: Brian Henn, Department of Civil and Environmental Engineering, University of Washington, Box 352700, Seattle, WA 98195-2700.

E-mail: bhenn@u.washington.edu

DOI: 10.1175/JHM-D-12-027.1

© 2013 American Meteorological Society
and spatial distribution via objective analysis (Liston and Elder 2006). In this type of application, techniques must be chosen to fill in missing data. While spatial interpolation via objective analysis to estimate missing data is common (e.g., Eischeid et al. 2000), there are many other methods for filling near-surface air temperature data.

Broadly, filling of missing temperature data is similar to the techniques of interpolation, extrapolation, and forecasting, in that available observations are used as predictors of missing data. Much effort has been devoted to interpolating temperature fields across the land surface at a given time. Techniques for spatial interpolation of air temperature include inverse-distance weighting (Daly et al. 2000) and thin-plate splines (Pape et al. 2009), as well as kriging (Tobin et al. 2011; Garen et al. 1994) and multiple regressions (Stahl et al. 2006). In particular, the well-established relationship between temperature and elevation, described by the lapse rate, allows for distribution of temperature across complex terrain, provided that the lapse rate can be accurately estimated (Minder et al. 2010; Rolland 2003; Daly et al. 2002; Dodson and Marks 1997). However, filling with an observed local lapse rate requires the presence of multiple observing stations, which is not always the case in the context of hydrologic modeling at the basin scale.

The above techniques rely on correlations in space to estimate unknown values at a given time, but in the case of missing temperature data, useful information is also contained within the portion of the time series that is not missing. This is because of the autocorrelation of the temperature time series, which presumably extends from known values both forward and backward into the missing period. For example, an autoregressive time series model can provide effective predictions at a single station by including the current temperature among the predictors of the temperature at the upcoming time step (Raible et al. 1999), and a common forecasting tool for temperature is “persistence.”

When the missing data are bounded in time by observations, forecasts and hindcasts in linear combination can be used to further improve estimation (Walton 1996). The methodology of Liston and Elder (2006) uses this approach with the explicit goal of filling in missing meteorological time series data. Other techniques, such as linear interpolation and polynomial fitting of missing data, have been found to be effective in reconstructing missing meteorological data (Claridge and Chen 2006). With these techniques, data from a single observing station are used to fill in missing periods in the station’s record.

The entire spatial and temporal correlation structure of a dataset can be described using empirical orthogonal functions (EOFs) (Von Storch and Zwiers 1999). EOFs are generated from a singular value decomposition of the dataset matrix, and they represent its dominant patterns of spatial and temporal variation (Preisendorfer 1988). The dataset can be represented as a linear orthogonal sum of the products of the spatial EOFs and their temporal weights at each time. Because the leading EOFs contain the bulk of the variance, they are more likely to represent broadscale patterns, while the latter EOFs likely represent local-scale patterns and instrument noise.

Because singular value decomposition cannot operate on matrices with missing data, Beckers and Rixen (2003) developed a method that iteratively estimates both the missing data and the complete EOFs. Filling is carried out by first inserting mean values in place of the missing data, then using a truncated series of EOFs to iteratively improve the estimates until convergence. The truncated series employs the leading EOFs to estimate missing data using the spatially and temporally coherent patterns while neglecting the local-scale noise associated with the subsequent EOFs. The EOF reconstruction approach has been used in oceanography, but EOF-based filling of missing data has not been widely applied to meteorological datasets such as air temperature.

Figure 1 summarizes the conceptual approaches of spatial and temporal filling of missing temperature data, including the above techniques. This study tests the methods’ skill by randomly removing available temperature data and then evaluating each technique’s ability to reconstruct the gaps. Skill is measured in terms of error and bias, as a function of number of stations and length of the missing period. We compared five techniques for filling missing near-surface air temperature: EOF reconstruction, time series interpolation as implemented in the MicroMet preprocessor (Liston and Elder 2006), spatial filling using an hourly lapse rate as well as a long-term lapse rate, and the combination of an hourly lapse rate with kriging.

To test the methods, we used five hourly near-surface air temperature datasets. Each set had between 19 and 63 stations and spanned at least 1 year, and all were located in areas of complex terrain. The largest was the National Oceanic and Atmospheric Administration (NOAA) Hydrometeorological Testbed (HMT) (Ralph et al. 2005) in the American River basin in the Sierra Nevada of California. The additional datasets were located in the French Pyrenees, the Yosemite region of the Sierra Nevada, the northern Cascades of Washington State, and the Colorado Rocky Mountains.

The goal of this study is to determine the optimal method for filling missing temperature data, given a
particular number of available temperature stations and gap characteristics. We are interested in the impact of the number of stations on the skill of each method as well as the impact of the length of the missing periods. With these findings, it will be possible to recommend a “best” method for practitioners looking to fill missing data in a dataset with particular characteristics.

Section 2 describes the air temperature datasets, while section 3 describes the different data filling methods. Section 4 presents the methods’ skill as a function of number of stations and length of gap. Section 5 explores the impact of the dataset characteristics on the methods’ performance and provides guidelines for the choice of filling method. Section 6 presents conclusions based on the results of the study.

2. Data

Five datasets of hourly near-surface air temperature were used as a basis for testing the skill of the filling methods. One was made up of previously unpublished distributed temperature data from the NOAA HMT-West watershed, the northern fork of the American River (basin size 885 km²). This area is located on the western slope of the Sierra Nevada, approximately along the I-80 corridor between Sacramento and Lake Tahoe. The other four datasets were from the eastern Pyrenees on the border of Spain and France (Pepin and Kidd 2006), Yosemite National Park, also in the Sierra Nevada of California (Lundquist et al. 2003), North Cascades National Park in Washington State (Minder et al. 2010), and Loch Vale in Rocky Mountain National Park, Colorado (Lundquist et al. 2008). We considered an additional dataset of mean daily temperatures from the Snowpack Telemetry (SNOTEL) network in Washington State and Oregon (Serreze et al. 1999; data subset from Raleigh and Lundquist 2012), where the stations were much less spatially dense. Figure 2 displays the location of the datasets. Table 1 lists the characteristics of each dataset, including the number of stations, period of record, and completeness after quality control.

The hourly temperature datasets were collected using several types of low-cost distributed instruments, such as Maxim iButtons and Onset Tidbits and HOBOs, as well as more standard meteorological stations where they were available. The SNOTEL dataset was collected with standard meteorological stations only. The northern Cascades dataset was linearly interpolated to an hourly time step from the original 90-min observations. The eastern Pyrenees data were aggregated to an hourly time series from 15-min observations. All data were quality controlled according to guidelines described by Meek and Hatfield (1994). Data that exceeded plausible maximum and minimum values were removed, as were points in which the rate of change between hours was deemed excessive, for example, >5°C h⁻¹. For the American River dataset, when daily air temperature ranges were ≥3.5°C and maximum daily temperatures were <5°C, sensors were considered snow covered, and these data were removed as well. For the eastern Pyrenees, Yosemite, northern Cascades, and Loch Vale datasets, only stations without apparent snow cover interference were considered.

After quality control, the completeness of the datasets varied substantially. When selecting for the years of the field campaign with the greatest temporal coverage (2008–10), the American River dataset was missing 16% of all station hours. The other datasets were more complete, ranging from the eastern Pyrenees (87% complete) to Loch Vale (99% complete). The SNOTEL

![Conceptual diagram for filling missing temperature data](image-url)
dataset had only 63% coverage because a number of entire station years had been removed in the quality control process.

The mean distance between stations in each dataset varied between less than 1 km (Loch Vale) and more than 20 km (American River), when calculated as the mean of all station pair distances in the dataset. Spatial correlations of temperature were significant at this scale. In fact, an examination of the American River dataset (not shown) revealed that spatial autocorrelation persisted even over the longest distances across the watershed, approximately 90 km. This is corroborated by the high degree of correlation (Table 1, mean correlation coefficients of +0.87 to +0.97) seen between stations within each dataset. It is likely that the synoptic features that control temperature, such as atmospheric ridges and troughs and the warm and cold sectors of frontal systems, were larger than the basins’ spatial scale.

**American River dataset sources**

The American River dataset was one result of a field campaign to make hydrometeorological observations across a watershed at a high spatial and temporal resolution. Three different data sources were combined to estimate near-surface air temperature across the basin.
of the American River, spanning water years (which begin on 1 October of the previous year) 2008–10. The California Department of Water Resources (CDWR) manages a network of 15 automated air temperature stations in or near the American River basin, 11 of which had sufficient temporal coverage. Data were obtained from the California Data Exchange Center (CDEC; http://cdec.water.ca.gov). These sites were supplemented by 10 additional weather stations with sufficient coverage deployed by the NOAA HMT (http://hmt.noaa.gov/). HMT sensors are mounted on a pole or mast approximately 3–4 m above ground level.

Additionally, to examine temperatures in locations where it would be impractical to install a traditional meteorological tower, we deployed 42 self-recording Maxim iButton temperature sensors, for a total of 63 sites. The iButtons were installed in trees, at heights of 2–4 m above the ground, depending on what was necessary to keep them above the snowpack, following the methods of Lundquist and Huggett (2008). The iButtons recorded temperature at either hourly or 2-hourly resolution. To maintain a consistent time resolution, the 2-hourly stations were linearly interpolated to hourly resolution before any further analysis.

The CDWR stations were missing 7% of the data, with a median gap length of 1 h, and the NOAA HMT stations were missing 8%, with a median gap length of 24 h. The iButton stations were missing 21% of the data, with a median gap length of 24 h. Much of the missing iButton data was due to the observation gaps during summer maintenance periods, as well as stations that were installed for only two of the three water years. Finally, winter storms caused simultaneous periods of missing data across all three data sources; for the iButtons this was due to snow cover, while for the other sources it was likely due to power outages and transmission failure.

3. Methods

Five techniques were compared in terms of skill in filling data gaps in near-surface air temperature data (Table 2). The EOF reconstruction technique followed the method of Beckers and Rixen (2003) and is referred to as the EOF method. The second method was an implementation of the MicroMet preprocessor (Liston and Elder 2006), herein referred to as MicroMet. Next, two types of hourly lapse rate–based filling were examined, one using a least squares fit and the other employing kriging interpolation. The final method considered was a long-term constant lapse rate.

The experiment was designed in the Monte Carlo style such that, for each iteration and dataset, a set of a varying number of stations was selected. Then, 5% of all of the available data in each set were removed in

<table>
<thead>
<tr>
<th>Method name</th>
<th>Reference</th>
<th>Basis</th>
<th>Lapse rate</th>
<th>Weighting</th>
</tr>
</thead>
<tbody>
<tr>
<td>A EOF reconstruction</td>
<td>Beckers and Rixen (2003)</td>
<td>Spatiotemporal</td>
<td>Computed implicitly</td>
<td>Computed</td>
</tr>
<tr>
<td>B MicroMet preprocessor</td>
<td>Liston and Elder (2006)</td>
<td>Temporal</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>C Hourly lapse rate</td>
<td>Stahl et al. (2006)</td>
<td>Spatial</td>
<td>Computed by regression</td>
<td>Least squares fit</td>
</tr>
<tr>
<td>D Hourly lapse rate + kriging</td>
<td>Stahl et al. (2006)</td>
<td>Spatial</td>
<td>Computed by regression</td>
<td>Kriging</td>
</tr>
<tr>
<td>E Long-term lapse rate</td>
<td>Stahl et al. (2006)</td>
<td>Spatial</td>
<td>Specified</td>
<td>Least squares fit</td>
</tr>
</tbody>
</table>

JUNE 2013 HENN ET AL. 933

Unauthenticated | Downloaded 03/25/22 08:50 AM UTC
a randomized fashion (see below). Each of the five filling methods to be compared was run on the dataset, generating five different filled versions. Each version was then compared to the original data in order to generate estimates of the root-mean-square error (RMSE) and bias at the hourly time scale. The hourly errors were sorted for further analysis by the number of available stations and the length of the period of missing data. Then the hourly results were also aggregated into daily temperature statistics ($T_{\text{min}}$ and $T_{\text{max}}$) in order to analyze the impact of filling on temperature at the daily time scale.

For each regional dataset, the number of stations considered in a given round was two, then four, then eight, etc., up to the total number of stations available for that set. A round consisted of 50 iterations where data were removed in a random, different pattern at each iteration and then filled using the different methods. From the iterations, the mean error for each method and its variance could be calculated.

a. Random data removal

At each iteration, a random set of stations was selected from the full dataset, and then portions of the stations’ records were removed. The data removal was random both in terms of its location within the dataset and the length of the period removed, which was allowed to vary widely (from 1 h to 1 yr of data), in order to test the impact of gap length on estimation accuracy. The gap locations and lengths were generated using the matrix laboratory (MATLAB; MathWorks, Inc. 2010) pseudorandom number function rand.m; in this way, a random assortment of gaps of different lengths was introduced into the dataset. This process removed 5% of the original data. Because the datasets already contained periods of missing data, this process served to add additional gaps to the record.

Figure 3 shows an example of data removal and the resulting filling estimates from the different methods. The number of hours removed is 3, 10, and 24 in Figs. 3a, 3b, and 3c, respectively, with the missing period beginning at midmorning local time in each case. The 3-h gap ended just before the observed maximum temperature, the 10-h gap included the maximum daily temperature but not the following minimum daily temperature, and the 24-h gap included both extremes. In this example, the mean error increased with gap length for the MicroMet and EOF methods but not for the lapse rate methods. It also illustrates the importance of the diurnal cycle in temporal interpolation of near-surface air temperature.

b. Filling method A: EOF reconstruction

A step-by-step description of the EOF reconstruction technique of Beckers and Rixen (2003) is provided in the appendix. This method reconstructs missing data using empirical orthogonal functions, derived from the original data. While EOFs in a complete dataset would typically be calculated using singular value decomposition (SVD.m in MATLAB), the presence of missing data requires an iterative approach. The method of Beckers and Rixen (2003) allows for the estimation of missing data.
values and full EOFs by first inserting mean values into the missing portions of the dataset and then calculating the EOFs. Because the resulting spatial EOFs and the time series of their magnitudes reconstruct the original data, a truncated version of the original dataset can be generated, using only as many EOFs as are deemed significant through validation. This provides an improved estimate of the missing information over simply inserting mean values, because the small-variance (i.e., noise) EOFs have been removed.

The selection of functions to include in the truncated set has been shown to influence the estimation error. Beckers and Rixen (2003) retained a portion of the dataset for validation and continued to add EOFs (ordered from most variance explained to least) to the reconstructed series until the validation error ceased to decline. We used a similar method: 2% of the data were retained for validation of the number of EOFs, and reconstruction was run using between 1 and 20 EOFs. The validation data were used to calculate the error as a function of the number of EOFs retained. The optimal number of EOFs was then selected for use in reconstruction. This validation process was repeated at every iteration so as to consistently use the optimal number of EOFs.

c. **Filling method B: MicroMet preprocessor**

The MicroMet preprocessor, described in Liston and Elder (2006), uses linear interpolation for 1-h gaps and the mean of the previous and next day’s temperature at a given hour to fill gaps of 2–24 h. At times this method was unable to fill in specific hours because of gaps at the same time of day on the previous or next day; here the missing periods remained unfilled. Longer gaps were filled using a periodic moving-average time series model that mimics the diurnal cycle of temperature. The moving-average model (Box and Jenkins 1976; Walton 1996) was provided with a training period that is equal in length to the missing period. A forecast was extended forward in time across the missing period, and an equivalent hindcast was extended backward in time from the period following the gap. The final estimation of the missing data was a weighted average of the two such that there were no discontinuities. These three steps of filling in MicroMet were hierarchical, such that shorter gaps were filled prior to longer ones, allowing for maximum coverage. This was a key difference between the MicroMet method and the others, because it allowed for adaptive estimation of the missing data depending on the length of the gap.

d. **Filling methods C and D: Hourly lapse rate and hourly lapse rate with kriging**

Spatial interpolation is a common technique for estimating missing temperature data, but because of the topographical influence in each of the datasets, the effect of elevation must be taken into account in any interpolation scheme. Lapse rate–based data filling uses a linear temperature–elevation relationship to estimate missing data at a given station on the basis of its elevation.

When implementing lapse rate–based filling, choices must be made both in terms of the lapse rate to be used and the weights that the available data points should be given (see Table 2). The lapse rate can be calculated instantaneously or held constant as a prescribed value. The weighting options include least squares fit and kriging as well as nearest-neighbor and inverse-distance methods. Inverse-distance and nearest-neighbor weighting were found to perform poorly and were not retained as filling methods. We calculated the instantaneous hourly lapse rate using a linear least squares fit to the available temperature and elevation data. This method provided the estimated lapse rate and sea level intercept and was applied if at least two stations were available for fitting. If only one station had data available at a given hour, then a constant, long-term lapse rate, calculated a priori from a temporal average of all station data, was used instead. If in any iteration, the stations were excessively close in elevation and the magnitude of the estimated lapse rate exceeded 20°C km\(^{-1}\) (due to dividing by a very small number to fit the slope), then the lapse method was not applied.

Because the lapse rate method may be applied from any base station, a weighting scheme also must be applied. The weighting techniques considered were a least squares fit to all available data (method C, hourly lapse rate) and kriging (method D, hourly lapse rate with kriging). In method C, the weights of the available stations were implicitly dictated by the least squares fit of the linear lapse rate; the missing data were estimated depending on where the station fell in the elevation–temperature regression. In this case the estimated temperature is given by

\[
T_{\text{est}} = T_0 + \Gamma z_{\text{est}},
\]

where \(T_0\) and \(\Gamma\) are the sea level intercept temperature and the lapse rate previously estimated from the least squares fit routine and \(z_{\text{est}}\) is the elevation of the target station.

In method D, the hourly lapse rate was combined with kriging weighting of the available stations. Kriging has been shown to effectively estimate missing data when the observed data have high spatial resolution (Gunes et al. 2006). The hourly lapse rate found as above was used to calculate the temperature at each station location if it were lapsed to sea level; the transformed temperature field was then the basis for kriging interpolation.
This is the technique of detrended kriging applied by Garen et al. (1994). From this detrended field, ordinary kriging (Webster and Oliver 2001), implemented using the Gstat package for MATLAB (www.gstat.org), was applied at each missing station location using the following steps. First, the sample semivariogram was estimated from the field, including its range, sill, and nugget; the semivariogram was assumed to be spherical. Then, the kriging equation selected the weights for the available stations, which minimized the variance of the estimate. The detrended temperature at the missing site was the weighted average of all the available sites’ detrended temperatures. Finally, the interpolated temperature at the missing site was lapsed back to its original elevation to provide the final estimate.

### e. Filling method E: Long-term lapse rate

The long-term lapse rate procedure assumed a constant long-term linear lapse rate, which was calculated on the basis of a least squares fit of the long-term temporal average temperature of all available stations for a given iteration. At a given hour, an additional least squares fit was used to establish the intercept of the linear temperature–elevation relationship \[ T_0 \] in Eq. (1) above, given the prescribed long-term slope. The missing data were estimated by Eq. (1), that is, by where a station’s elevation fell on the elevation–temperature regression. Note that methods C, D, and E become identical for the case in which only one station has data available at a given hour, because in that case a prescribed lapse rate is the only feasible method for estimating missing temperatures.

### f. SNOTEL dataset comparison

To examine the effect of station density on the results, we also compared the results from the SNOTEL dataset with the results from the five denser networks. Because the SNOTEL data is at daily temporal resolution, we first calculated the average daily temperature from the hourly data at each of the dense-network sites. Days with more than 6 h of missing data were coded as missing. The same type of Monte Carlo–style validation analysis was then conducted on the six daily datasets. All methods were tested with the exception of MicroMet, which must run at hourly resolution.

### g. Performance metrics

The methods’ skills in filling missing data were evaluated in several ways. First, the hourly accuracy of the filled gap periods was calculated for each method. The accuracy was evaluated in terms of RMSE and bias compared to the original dataset. RMSE was chosen as the primary metric over the mean absolute error (MAE). Both give relevant information regarding performance, and an analysis of our results found that the magnitude of MAE was typically about 75% of that of RMSE for the filling methods. Whether minimizing MAE or RMSE is more important is likely to depend on the specific application.

Time periods in which data were missing in the original dataset were excluded from evaluation. At the hourly time scale, the error was calculated over all hours in which data had been removed, not including “natural” gaps. Performance was evaluated in terms of the number of stations available as well as the length of the gap, in order to determine whether each technique’s performance was dependent on these characteristics.

The reconstructed datasets were also used to calculate daily temperature statistics \( T_{\text{min}} \) and \( T_{\text{max}} \), which were compared against the equivalent values derived from the measured data. The daily statistics were used because they are frequently utilized as input to hydrologic models. Days in which the removal of data did not change \( T_{\text{min}} \) and \( T_{\text{max}} \) (because those values were not removed) were not considered when assessing the methods’ skills.

A two-tailed Student’s \( t \) test was used to determine statistical significance of differences between the methods. The lowest RMSE for a given number of stations and gap length was tested against the other methods’ RMSE to determine if this outcome was statistically significant at 95% confidence. The 50 Monte Carlo–style iterations provided the sample population for the \( t \) test.

### 4. Results

For each dataset, the number of stations was varied from two to the maximum available, and the methods’ performances were evaluated. To examine how the length of the gap affected method skill, the mean hourly errors were sorted by the length of the gap in which they were generated. The resulting RMSE from the American River dataset, spanning the two independent variables (number of stations and gap length), is plotted in Fig. 4. The error from the EOF method (method A) was strongly dependent on the number of available stations (Fig. 4a). The method performed well (<2°C RMSE) when more than eight stations were available, while it generated errors >5°C RMSE when only two stations were available. The EOF method showed a very limited dependence on the length of the missing period; only when missing periods approached 1 yr (not shown) did the EOF performance begin to degrade.

In contrast, the MicroMet preprocessor (method B) exhibited strong dependence on the length of the data gap (Fig. 4b). This method applied linear interpolation
to 1-h gaps, and the resulting errors were small (eastern Pyrenees: 0.62°C; Yosemite: 1.08°C; Loch Vale: 0.97°C) for the datasets in which linear interpolation was not applied to generate hourly time series. This confirmed that its use in the other datasets did not introduce excessive errors. The MicroMet method also performed very well when the gap lengths were less than 6 h, in which case the gaps were filled by averaging the previous and following day’s temperatures. However, the method deteriorated when the gaps were longer than 1–3 days. Finally, the MicroMet method was insensitive to the number of available stations, as it used only data at the given site to generate estimated temperatures.

The hourly lapse rate (method C) and the hourly lapse rate plus kriging methods (method D) generated results that were similar to each other (Figs. 4c,d). Both methods were dependent on the number of available stations, with best performance for cases with more than eight stations. The addition of kriging slightly improved the estimation accuracy when at least 32 stations were used. Finally, the long-term lapse rate (method E; Fig. 4e) generated slightly smaller errors relative to the other lapse rate techniques when few (4–8) stations were available. All of the lapse rate methods were sensitive to the number of stations but not to the length of the gap, even with periods of missing data approaching the length of the record.

Results for the five hourly datasets (American River, eastern Pyrenees, Yosemite, northern Cascades, and Loch Vale) are shown in Fig. 5. Here the error is given only in terms of the number of available stations, for an average of all gap lengths of 1–24 h. Thus, Fig. 5 emphasizes the effect of the number of available stations, which is important for the EOF and lapse rate–based method, but not the effect of gap length, which is important for the MicroMet method (MicroMet generated mean RMSE greater than 7°C for gaps longer than 3 days, for example). The other datasets followed the same general patterns as the American River dataset. However, the EOF and lapse rate–based methods performed relatively better on the eastern Pyrenees and Loch Vale data than on the Yosemite, northern Cascades, or American River data. These differences in performance were explored in more detail and are discussed in section 5a below.

In all datasets, the EOF and three lapse rate methods were sensitive to the number of stations that were available. The EOF method and lapse rate–based methods generally produced errors that were smaller than those
of the MicroMet preprocessor, so long as there were at least 4–16 observing stations (depending on the dataset).

No method was found to have a statistically significant warm or cold bias at the hourly time scale. The mean biases also became very small, with magnitudes \(0.18\) °C, as the number of stations increased beyond eight. Because it appeared that all of the methods were unbiased, the determinant of whether a method was the most effective for hourly simulations was simply the RMSE.

5. Analysis and discussion

a. Method sensitivity to dataset physical characteristics

To better understand the results shown above, we examined the correlation between the characteristics of the dataset and the methods' performance in the Monte Carlo simulations. For each dataset and each simulation run, we recorded the following characteristics about the randomly selected set of stations: the mean vertical separation between stations, the mean horizontal separation between stations, and the mean temporal correlation between stations. Each run's RMSEs for the EOF, hourly lapse, and long-term lapse methods were also recorded, and then the Pearson product-moment correlation coefficient was calculated between the error and each of the three characteristics. Only iterations with two or four stations (100 total for each dataset), and not the larger station sets, were included in the correlation analysis. With the larger number of stations, the mean characteristics varied little between Monte Carlo runs, and so it was not possible to resolve the error correlation. The MicroMet method was not included in the correlation analysis as it is not dependent on the relationship between stations; the hourly lapse rate with kriging method was also excluded as its results were quite similar to the hourly lapse rate method.

The correlation analysis showed that the vertical separation between stations was a predictor of the methods' accuracy (Fig. 6a). The strongest correlations occurred between RMSE and the mean vertical station separation, with lower error associated with smaller separation. This was true for the lapse rate methods for
all datasets and for the EOF method for all datasets except Yosemite and Loch Vale. The horizontal separation of stations (Fig. 6b) was not as significantly correlated with the methods’ error as the vertical separation. Additionally, significant negative associations were found between the stations’ correlations to each other and RMSE (Fig. 6c). In other words, station sets with greater temporal correlation produced smaller estimation errors.

Examining the relative physical characteristics of the five datasets and the method sensitivities to them explained many of the results seen in Fig. 5. We plotted the distributions of the physical characteristics for each dataset from the Monte Carlo simulations (Fig. 7). The Yosemite dataset, for example, had much greater vertical range than the other datasets. Given that vertical distance was correlated with error (Fig. 6a), this may have explained the relatively poor performance of the lapse rate and EOF methods in the Yosemite, American River, and northern Cascades datasets. Similarly, these methods performed better in the eastern Pyrenees and Loch Vale datasets, which had relatively small vertical separation between stations. Additionally, relatively low correlation between stations may be another reason for the weaker performance of the lapse rate methods on the northern Cascades dataset, as station correlation was also found to be a significant predictor of accuracy for the EOF and lapse rate methods (Fig. 6c).
Error in predicting the daily minimum and maximum temperatures ($T_{\text{min}}$ and $T_{\text{max}}$) gave some indication of variation of the error according to the time of day across the datasets. In general, the equivalent RMSE plots for $T_{\text{min}}$ and $T_{\text{max}}$ (not shown) are qualitatively similar to Fig. 5: the Yosemite, northern Cascades, and American River datasets produced higher error than the Loch Vale and eastern Pyrenees datasets for a given method. In contrast, the differences between the daily minima–maxima and hourly errors are plotted in Fig. 8, which gives an indication of the relative difficulty of filling daily extremes. Relative to the hourly error, $T_{\text{min}}$ fared worse at Yosemite, while $T_{\text{min}}$ and $T_{\text{max}}$ fared better, relative to the hourly error, for the northern Cascades dataset.

Station density was found to have only a small effect on method performance. Because all five of the datasets are relatively dense, we compared them with temperature data from the SNOTEL network of meteorological stations located in the western United States. This dataset of daily mean temperatures comprises 107 stations in Washington and Oregon, across a distance of more than 700 km. We ran the EOF reconstruction, hourly lapse rate, hourly lapse rate with kriging, and long-term lapse rate methods (A, C, D, and E) on this dataset. Because it is a daily dataset, we ran the methods on daily mean temperature versions of the other five datasets and compared the relative performance of the methods across the different spatial scales. The gaps in this analysis ranged from 1 day to near the length of the whole dataset; the results for these methods were not found to be sensitive to gap length. The MicroMet method was not included in this analysis because it is intended to run only at an hourly time step.

The methods’ performance on the SNOTEL dataset, relative to the other sites, is shown in Fig. 9. For the SNOTEL dataset (Fig. 9a), the error in the mean daily temperatures declined for all methods as the number of stations increased, to approximately 2°C for station sets of 32 or greater. It is apparent that the smaller, denser datasets (Figs. 9b–f) generated somewhat lower errors than the larger SNOTEL dataset, with mean daily temperature errors approaching 1°C for the denser datasets. However, the SNOTEL mean daily temperature error was of a similar magnitude (about 2°C) to the denser dataset errors, despite covering a much larger spatial range.

We found that the performance of the methods varied seasonally, but not in a consistent fashion between datasets. We compared the RMSE of each method between the four seasons [defined here as December–February (DJF), March–May (MAM), June–August (JJA), and September–November (SON)] at each of the five datasets at the hourly time scale. Some statistically significant seasonal differences were found. The MicroMet method produced a greater error in winter in the American River dataset, by approximately 1°C over the other three seasons, and the lapse rate methods had greater error in the spring and summer in the northern Cascades dataset. However, we did not find consistent seasonal differences between the methods’ performances across the five datasets.

b. EOF filling of long-term missing data

Because the EOF reconstruction technique uses both spatial and temporal correlations to estimate missing data, it was able to fill very long gaps with better skill than the lapse rate–based methods if sufficient stations

![Fig. 8](image_url)
were available (the upper right region of Fig. 4a). We also examined the case in which the missing data are not gaps but stations that have been removed entirely, for example, after the conclusion of a temporary field campaign. If there are permanent stations available, they can be used to “train” the EOF reconstruction to fill the missing period indefinitely. We tested the accuracy of the EOF method in this situation compared with the lapse rate approaches by selecting two target stations along with multiple permanent stations from the American River dataset. The target stations had 1 year of available data and two missing (removed) years; the permanent stations had 3 years of fairly continuous (at least 90%) coverage. We then applied the EOF, hourly lapse rate, and long-term lapse rate methods and calculated the error in estimating the target stations’ missing years.

In this analysis (Fig. 10), the RMSE of the EOF method was lower than that of the two lapse rate methods if there was only one permanent station or if there were more than six such stations. The long-term lapse rate had lower error than the hourly lapse rate method at all numbers of stations and the lowest error overall in the case of two permanent stations.

c. Best practices for filling temperature data

On the basis of the above results, it was apparent that certain methods for filling missing temperature data performed better or worse depending on the number of stations and the length of the period of missing data. Figure 11 shows the situations (as a function of gap length and number of stations) in which one method had the lowest RMSE in at least two of the five datasets. Additionally, if the RMSE of the most accurate method had statistical significance (at the 95% level according to the t test) in at least three of the five datasets, then it is designated with an × symbol.
Some broad patterns became evident when the results were plotted in this manner. First, the MicroMet preprocessor was the most accurate method, with statistical significance, for single-hour gaps regardless of the number of stations. This demonstrated the low error associated with the linear interpolation of 1-h gaps. MicroMet was also the best method when only two stations were available for gaps of 1 h to 3 days, confirming that the diurnal cycle assumption used in filling these gaps was also fairly effective. For gaps longer than 24 h, however, spatial or spatiotemporal methods were more accurate, which demonstrated the importance of the gap length in time series interpolation.

Second, the EOF technique is most accurate, with statistical significance, when there were 16 or more stations, for gaps of 2 h to 14 days. This demonstrated the importance of the number of stations when applying EOF reconstruction, which was also the most accurate method when there were only two stations but the period of missing data was 3 days or greater. The MicroMet approach had the lowest accuracy of any method for all gaps of 6 days or longer, with statistical significance. This showed that very long gaps were not easily filled with time series interpolation. In the middle of the figure lies a region of lower certainty of the best method, in which various lapse rate methods appeared to be most accurate. For example, the long-term lapse rate method was most accurate for four stations for gaps of 1–30 days. However, in no case did any of the lapse rate methods have statistically significantly greater accuracy than the EOF method in at least three of the datasets.

6. Conclusions

We carried out a Monte Carlo–style validation study of five near-surface air temperature filling techniques, using five spatially dense hourly observational datasets from western North America and Europe. When choosing a method with which to fill short-term gaps in observed
temperature datasets, the study indicated that while bias is not likely to be a systemic concern, the method with the lowest error will depend strongly on the length of the gap and the number of stations available. When gaps were short (less than 24 h), MicroMet was the most accurate method, so long as the number of available stations was not large (generally less than 16). It had the additional advantage of requiring only a single station in order to fill in data. Short gaps were easier to fill because time series interpolation could be applied accurately. However, when at least 16 stations were available, or when the gaps exceeded 24 h, the EOF method generally became the most accurate.

The sensitivity of the MicroMet method to the length of the period of missing data is apparent from the gradient of error with the length of gap in Fig. 4b. This illustrated that the assumption of a diurnal cycle allowed for accurate estimates in temporal interpolation for subdaily gaps. However, longer gaps were more difficult to fill; this was presumably because additional factors, such as synoptic variability, came into play.

Hourly lapse rate–based filling displayed more variable performance across the datasets. This may have been because of the potential for deviation from a linear lapse rate in complex terrain under some synoptic conditions, for example, the effect of cold air pooling on \( T_{\text{min}} \) in the Yosemite dataset (Lundquist and Cayan 2007; Lundquist et al. 2008) or the effect of low solar radiation in limiting \( T_{\text{min}} \) and \( T_{\text{max}} \) errors in the northern Cascades. The accuracies of the lapse rate and EOF methods were found to be sensitive to the vertical separation of the stations, with larger separation resulting in greater error. Combining kriging interpolation with the hourly lapse rate method offered modest improvements over a least squares fit.

With regard to the horizontal separation between stations, we found that the EOF and lapse rate methods’ performance declined only slightly when the spatial scale of the dataset was much larger (Fig. 9). For the large-scale dataset, daily mean temperature errors were generally within 1°C of those for denser datasets. This indicated that practitioners may be able to make use of publicly available station data that are several hundred kilometers from the study site in order to attain a large number of reference stations. The inclusion of additional stations allows for more accurate estimation of gaps using EOF or lapse rate methods, as we have shown. The results also indicated that finescale variation in near-surface air temperature may be as large as variation over much longer distances.

For practitioners, Fig. 11 offers guidelines on how to approach filling of missing air temperature data. In the case when observations have ceased at a site, the EOF method may offer the ability to reconstruct long stretches of missing data (Fig. 10). The lower error of the EOF method relative to the lapse rate methods was presumably attained by utilizing the training period. The results also suggested that for a given dataset, a hybrid approach might be applied in which missing gaps are estimated with different methods depending on their length and the number of available stations.

Finally, for all methods except MicroMet, RMSE declined with increasing numbers of observing stations. This highlighted the value of distributed temperature measurements in watershed-scale applications, particularly given that they can be made with accuracy and at relatively low cost.

Acknowledgments. We thank Courtney Moore for her work on the American River dataset and Nic Wayand, Shara Feld, Kael Martin, Susan Dickerson, Karl Lapo, and Nicoleta Cristea for technical support and review. We would like to acknowledge those who allowed us to use the additional datasets: Nick Pepin and David Kidd for the eastern Pyrenees dataset; Caitlin Rochford and David Clow for their work on the Loch Vale dataset; Natalie Low for her work on the northern Cascades dataset; and Dan Cayan, Mike Dettinger, Heidi Roop, Brian Hugett, and Jim Roche for their work on the Yosemite dataset. NOAA provided funding for the American River dataset through its Hydrometeorological Testbed, and this publication was partially funded by the Joint Institute for the Study of the Atmosphere and Ocean (JISAO) under NOAA Cooperative Agreement NA10OAR4320148, Contribution 2023. We would like to recognize Andrey Shcherbina for work on the implementation of the EOF-based data reconstruction algorithm. B.H. recognizes support from a National Defense Science and Engineering Graduate Fellowship. J.D.L. and A.F. recognize support from NSF Grant EAR-0838166. M.R. was partially supported by NASA Headquarters under the NASA Earth and Space Science Fellowship Program Grant NNX09AO22H.

APPENDIX

EOF Reconstruction Algorithm

Following Beckers and Rixen (2003), we implemented the EOF reconstruction method using the following pseudocode algorithm. The function receives \( n \times m \) data matrix \( \mathbf{X} \) with missing data and overall mean equal to zero.

1. \( \mathbf{X}' = \mathbf{X} \) (good); good = 98% of all available data retained for validation; val = 2% removed.
2. \( \mathbf{X}(\text{missing}) = 0 \); replace missing values with unbiased guess (0).
FOR $N$ (number of EOFs) = 1–20, or number of spatial sites $m$ if less than 20.

FOR iteration = 1 to convergence.
3. $[U, D, V] = \text{svd}(X')$; perform singular value decomposition.
4. $U_r = U(:, 1:N)$; truncate components using first $N$ EOFs.
5. $D_r = D(1\!:N, 1\!:N)$.
6. $V_r = V(:, 1:N)$.
7. $X_r = U_rD_rV_r'$; reconstruct estimated matrix $X_r$.
8. $X_r(\text{good}) = X(\text{good})$; restore original data at all locations except where missing.
9. var $X_r = \sum[(X_r - X)^2]$; calculate variance of estimate from initial guess.
10. var $X = \sum[(X')^2]$; calculate variance of entire estimate.

IF var $X_r$/var $X <$ tolerance; test for convergence by comparing variances.
11. BREAK; go to outer FOR loop.
END
END

12. error($N$) = sum([($X_r(\text{val}) - X(\text{val}))^2$]); calculate squared error of estimate for $N$ EOFs.

END
13. NOPT = find(error == min(error)); optimal number of EOFs is that which minimizes error.

Once the optimal number of EOFs was established, we repeated the inner FOR loop to calculate the spatial and temporal weights and variances using NOPT. Then, these spatial and temporal functions and the variances were returned to the main routine, and the estimated missing values could be obtained by applying line 7 above.

REFERENCES


W00D04, doi:10.1029/2008WR00979.


