# Inherent Bounds on Forecast Accuracy due to Observation Uncertainty Caused by Temporal Sampling

MARION P. MITTERMAIER

*Numerical Modelling, Weather Science, Met Office, Exeter, United Kingdom*

DAVID B. STEPHENSON

*Exeter Climate Systems, Department of Mathematics and Computer Science, Exeter University, Exeter, United Kingdom*

(Manuscript received 29 April 2015, in final form 23 July 2015)

ABSTRACT

Synoptic observations are often treated as error-free representations of the true state of the real world. For example, when observations are used to verify numerical weather prediction (NWP) forecasts, forecast–observation differences (the total error) are often entirely attributed to forecast inaccuracy. Such simplification is no longer justifiable for short-lead forecasts made with increasingly accurate higher-resolution models. For example, at least 25% of $t + 6$ h individual Met Office site-specific (postprocessed) temperature forecasts now typically have total errors of less than 0.2 K, which are comparable to typical instrument measurement errors of around 0.1 K. In addition to instrument errors, uncertainty is introduced by measurements not being taken concurrently with the forecasts. For example, synoptic temperature observations in the United Kingdom are typically taken 10 min before the hour, whereas forecasts are generally extracted as instantaneous values on the hour. This study develops a simple yet robust statistical modeling procedure for assessing how serially correlated subhourly variations limit the forecast accuracy that can be achieved. The methodology is demonstrated by application to synoptic temperature observations sampled every minute at several locations around the United Kingdom. Results show that subhourly variations lead to sizeable forecast errors of 0.16–0.44 K for observations taken 10 min before the forecast issue time. The magnitude of this error depends on spatial location and the annual cycle, with the greater errors occurring in the warmer seasons and at inland sites. This important source of uncertainty consists of a bias due to the diurnal cycle, plus irreducible uncertainty due to unpredictable subhourly variations that fundamentally limit forecast accuracy.

## 1. Introduction

Observations are used in several ways in numerical weather prediction (NWP): for defining initial conditions, verification, and postprocessing. For a long time the importance of surface observations seemed to be waning, mainly due to the sparseness of observing locations compared to the spatial coverage that radar or a satellite can provide. Yet for verification and postprocessing of surface variables there is no substitute.

Quality control of surface observations is a hugely important and time-consuming task, considering the efforts that go into maintaining temperature records for climate change purposes (Morice et al. 2012; Hansen et al. 2010; Smith et al. 2008). For weather forecasting the quality and accuracy of observations is also important. With improving horizontal resolution of short-range forecast models, often to convection-permitting kilometer scale, the expectation is that forecast errors will continue to reduce and skill will increase. While instrument errors for some atmospheric variables such as temperature are ~0.1 K [(World Meteorological Organization) WMO 2008] the impact of location and altitude of an observing site compared to the model gridbox representation of the orography could be a large component of any total error, which may well be larger than the instrument error. The effect of temporal sampling has generally also not been considered.

Observations are most often treated as absolute truth (i.e., they are considered to be representative of the true state), so if the model does not match the observed value, the forecast is assumed to be wrong. For example,

*Corresponding author address:* Marion P. Mittermaier, Met Office, FitzRoy Rd., Exeter, EX1 3PB, United Kingdom.
E-mail: marion.mittermaier@metoffice.gov.uk

DOI: 10.1175/MWR-D-15-0173.1

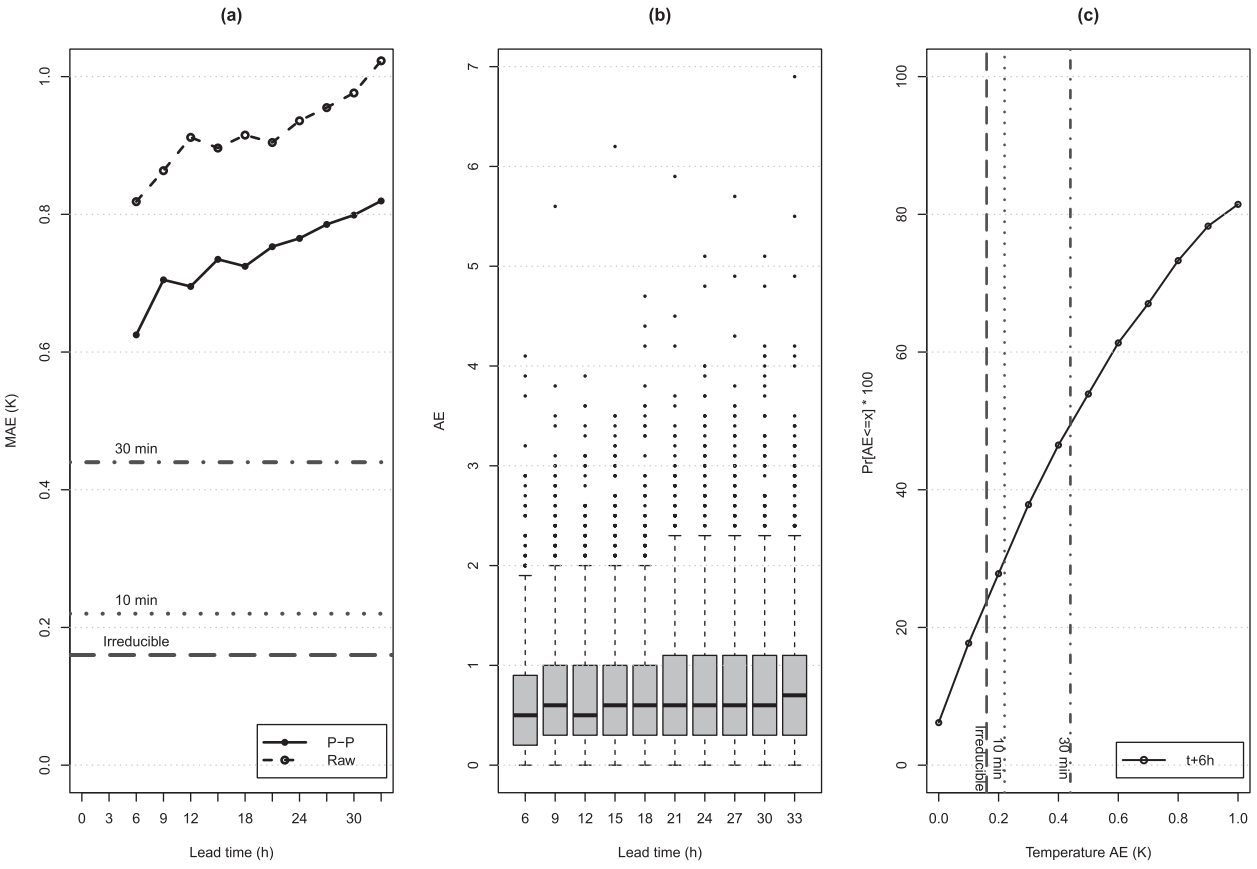


FIG. 1. (a) Annual MAE of raw and postprocessed (p-p) temperature forecasts for London Heathrow (03772) for 2014 at a range of lead times (in h). (b) Distribution of postprocessed absolute errors (AE) as a function of lead time showing outliers. (c) Cumulative distribution function of postprocessed AE for $t + 6$ h. The horizontal lines in (a) and vertical lines in (c) indicate the MAE when hourly observations expected for a perfect forecasting system (see section 4).

Fig. 1 provides 2014 verification statistics for raw and postprocessed temperature forecasts for Heathrow (03772) at lead times between $t + 6$ and $t + 33$ h. These are verified against hourly synoptic observations. The postprocessed forecasts are based on a lagging and blending of the Kalman filtered (KF) raw 1.5-km Met Office Unified Model (UM) site-specific output. The annual mean absolute error (MAE) in Fig. 1a is between 0.6 and 0.8 K for the postprocessed forecasts and between 0.8 and 1 K for the raw forecasts. The distribution of 3-hourly absolute errors for the postprocessed forecasts in Fig. 1b shows that there are some large values though most are less than 1 K. Considering the absolute errors in Fig. 1b as a cumulative distribution function, Fig. 1c shows that at $t + 6$ h around 40% of absolute errors are less than 0.5 K with ~5% of forecasts considered "perfect," and ~15% of forecasts with absolute errors less than or equal to 0.1 K, the instrument measurement error. So even now the interpretation of "forecast error" is compromised 20% of the time, by equaling (or ignoring) instrument measurement limits.

Only a few published studies have investigated the impact of observation error, as commented on by Jolliffe and Stephenson (2011) in the concluding chapter, and none have explicitly considered uncertainties introduced as a result of temporal sampling. Saetra et al. (2004) explored the impact of observation error through the addition of normally distributed noise to the true state. Bowler (2006) proposed a method for contingency table–based metrics. The method assumes that observations are not correlated in space or time. Bowler (2006) argues that a verification metric should not be affected by the quality of the observations network (i.e., given a perfect forecast, the use of an erroneous observation should still yield a perfect forecast). He argues that an approach such as that proposed by Candille and Talagrand (2005) describing the observation error as a probability density function will penalize a perfect forecast. Bowler (2008) subsequently used data assimilation–derived covariance estimates of the observations error to randomly perturb individual ensemble members. Santos and Ghelli (2012) extended the approach by Candille and Talagrand (2008) who

TABLE 1. Summary of synoptic observing times for a range of different atmospheric variables, with the UM equivalent, where HH stands for the hour and the value following it refers to the minutes before the hour HH. Note that the observing practice may differ in other countries.

| Variable | SYNOP | UM |
|---|---|---|
| Temperature | Instantaneous at HH-10 | Instantaneous time step nearest the hour |
| Wind (speed and direction) | 10-min average between HH-20 and HH-10 | Instantaneous time step nearest the hour |
| Cloud-base height and total cloud amount | Manual: instantaneous at HH-10 | Instantaneous time step nearest the hour |
| | Automated: exponential aggregate over 40 min between HH-50 and HH-10 | |
| Visibility | 1-min sample at HH-10 | Instantaneous time step nearest the hour |
| Precipitation | Accumulation (for hourly between HH-70 to HH-10) | Accumulation between time steps nearest HH and HH-60 |

considered empirical distributions to provide a measure of the spatial "representativeness" error. Koh et al. (2012) also considered the temporal scaling at a point through the use of spectral analysis. Röpnack et al. (2013) proposed a probabilistic approach based on Bayes's theorem.

This study investigates the potential contribution that temporal sampling uncertainty (through mismatches and subhourly variability) can make to forecast errors. We address the following questions:

(i) How can we best characterize the subhourly observational variability as a function of time of year and location?
(ii) What is the impact of this subhourly observational variability on verification statistics?

Section 2 provides an overview of the observations used for this study. Section 3 describes the approach taken in deriving subhourly variability. The impact of this subhourly variability on verification metrics is discussed in section 4. Conclusions follow in section 5.

## 2. Data

Real-time hourly synoptic observations are exchanged via the WMO's Global Telecommunication System (GTS). The WMO guide on observations (WMO 2008) defines "the representativeness of an observation is the degree to which it accurately describes the value of the variable needed for a specific purpose." It goes on to say "synoptic observations should typically be representative of an area up to 100 km around the station, but for small-scale or local applications the considered area may have dimensions of 10 km or less." The concept of representativeness described here is hard to reconcile with kilometer-scale NWP modeling, which can show considerable detail and variability at spatial scales less than 10 km, at least in part through the use of more detailed orography, which has clear impacts on temperature, fog, low cloud, and winds, to name but a few. Furthermore, for verification it is also generally assumed that the observation is temporally representative of the hour that it was reported.

### a. When are synoptic observations taken?

Table 1 provides a summary of when hourly synoptic observations are taken and the typical model equivalent, based on UM output protocol. The WMO recommendation is that synoptic observations not be taken more than 10 min before the hour (WMO 2014). In the United Kingdom, most observations are taken 10 min before the hour, though observing practices may differ in other countries. Temperature is truncated and reported to one decimal place. Note that for the UM the output is always the nearest model time step to the hour, except for precipitation. Time steps are different for each model configuration; for example, the current time step for the 1.5-km deterministic model (UKV) is 50 s, while the 2.2-km Met Office Global and Regional Ensemble Prediction System (MOGREPS-U.K.) uses a 75-s time step. The global model (GM) currently runs with a 10-min time step. Clearly there is at least one time step mismatch between when the model produces hourly output, and when the observation is taken. In some instances (e.g., wind and automated cloud parameters) aggregates are compared to instantaneous model output.

### b. What do subhourly observations look like?

The Met Office has access to 1-min observations from the national observing network. From these the hourly synoptic observations are reported in the surface synoptic observations (SYNOP) message that is transmitted worldwide via the GTS. Figure 2a shows the 1-min temperature time series for June 2013 at Heathrow (03772). The hourly SYNOP observations are superimposed. Although it is often assumed that temperature is a smoothly varying time series process, closer inspection of a smaller section of the time series in Fig. 2b shows that there are also fast irregular fluctuations. These rapid fluctuations may be due to the passage of fronts, changes in cloudiness, or turbulent mixing.

For this study a small sample of U.K. synoptic stations was selected to reflect a number of different geographical locations: upland, inland, and coastal (see Fig. 5 for a map). Eskdalemuir (03162) represents an upland site.

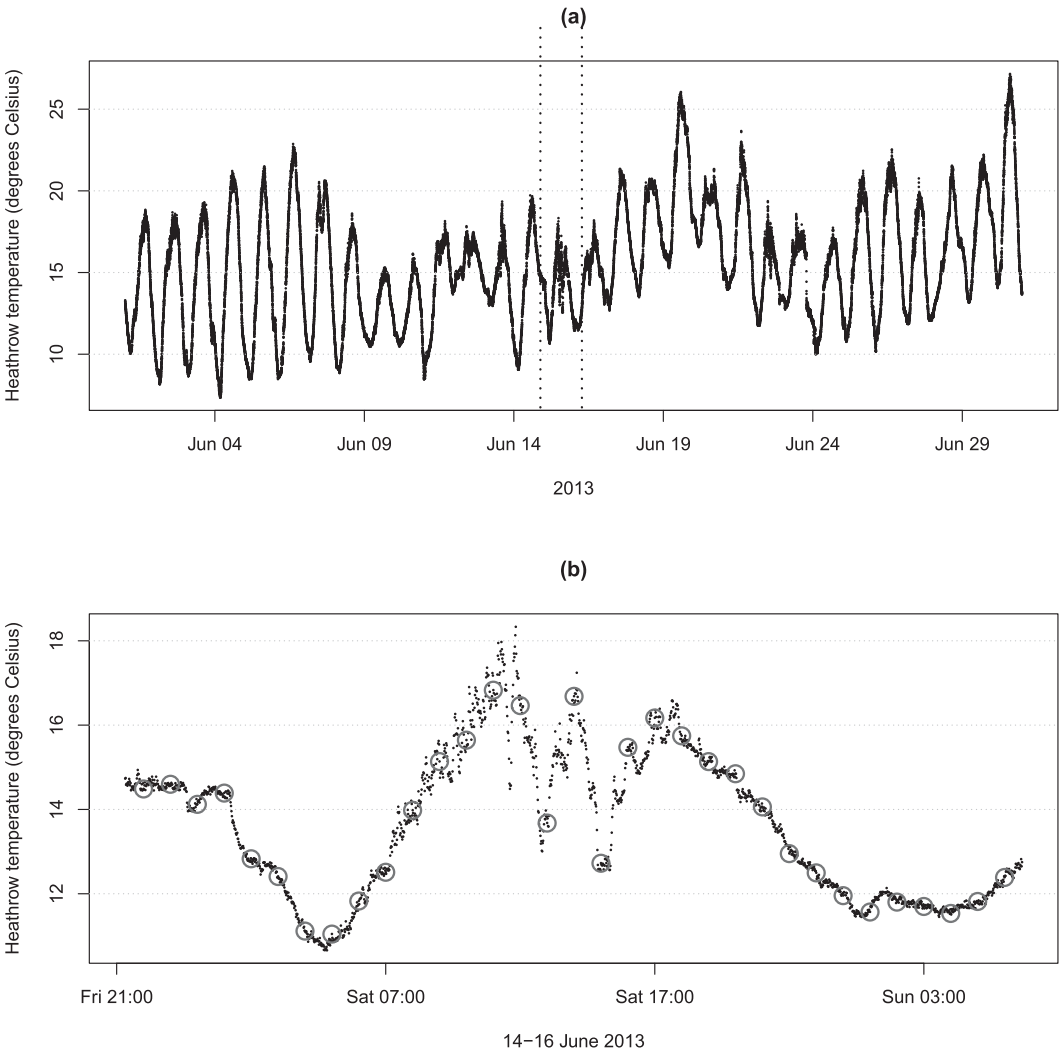


FIG. 2. Monthly time series of temperature from Heathrow (03488) for June 2013. (a) The 1-min series; (b) zoomed time series showing 2200 UTC 14 Jun–0600 UTC 16 Jun. Gray circles show the hourly synoptic observations, highlighting large subhourly variations.

Benson (03658) and Heathrow (03772) represent typical inland ''rural'' and ''urban'' sites, respectively. South Uist (03023) in northwest Scotland, St. Athan (03716) on the south coast of Wales, St. Mary's (03803) on the Scilly Isles off southwest England, and Weybourne (03488) on the North Sea coast, represent a range of maritime exposures.

## 3. Modeling approach

Our focus is on observation uncertainty caused by forecasts being extracted $m > 0$ minutes later than measurements that are taken every hour. Meteorological variables such as temperature in particular can exhibit strong diurnal trends and be serially autocorrelated. Any proposed method must account for such behavior, and be applicable to a range of different variables, geographical locations, and seasons. It is therefore of interest to develop a model for the observation increment $Y_{d,h,m} = X_{d,h,m} - X_{d,h,0}$, where $X_{d,h,m}$ is a random variable representing a measurable observable taken on day $d = 1, 2, \ldots, D$, hour $h = 0, 1, \ldots, 23$, and minute $m = 0, 1, \ldots, 59$.

The increment can be interpreted as the optimal forecast error that could be obtained if one were able to issue *perfect* forecasts of observations $X_{d,h,m}$ that are then verified against observations $X_{d,h,0}$ measured $m$ minutes earlier.[1] Therefore, the sample mean of $|Y_{d,h,m}|$ provides an

[1] Note that this hypothetical best-case scenario would require perfect weather forecasts and no unpredictable measurement errors.

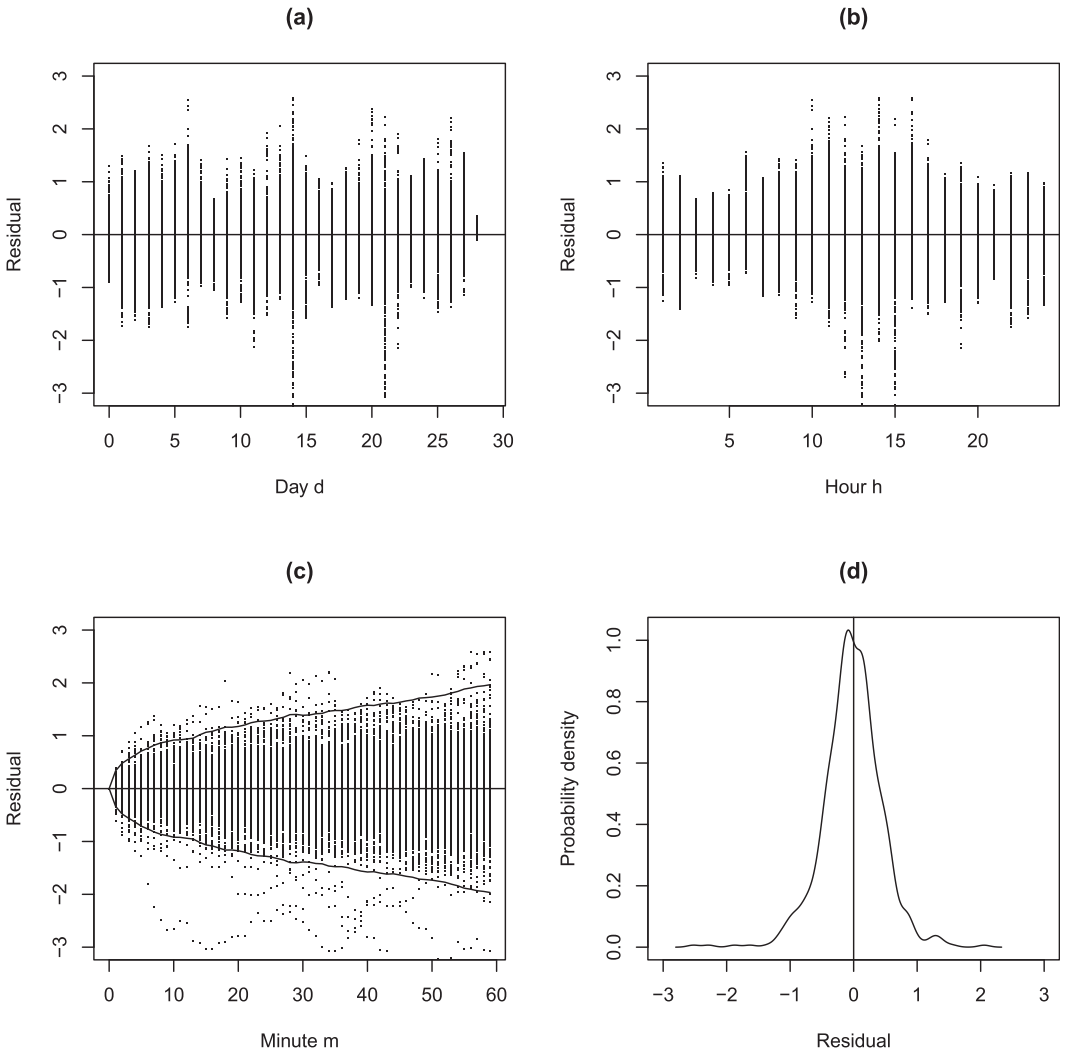


FIG. 3. Residual diagnostics for the model fit to June 2013 at Heathrow. (a) Checking whether the residuals $\varepsilon_{d,h,m}$ depend on day $d$ and (b) checking whether the residuals depend on hour $h$ of the day. (c) Standard deviation of $\varepsilon_{d,h,m}$ as a function of minute $m$ from the hour with $\pm$ 3 standard deviations and (d) $\varepsilon_{d,h,m}$ for $m$ = 30-min offset.

estimate of the smallest MAE that could ever be achieved (i.e., by issuing perfect forecasts of the observations).

The increment may be considered to be the sum of a diurnally varying trend component and a random noise component:

$$Y_{d,h,m} = \mu_{h,m} + \varepsilon_{d,h,m}. \quad (1)$$

For simplicity, we shall assume that the trend is linear throughout each hour $\mu_{h,m} = \alpha_h m$ with slope parameters $(\alpha_0, \ldots, \alpha_{23})$ that depend on the hour in the day. The noise $\varepsilon_{d,h,m}$ is assumed to be a random variable having an expectation of zero and a variance that depends only on the minute within the hour [i.e., $E(\varepsilon_{d,h,m}) = 0$ and $\mathrm{Var}(\varepsilon_{d,h,m}) = \sigma_m^2$]. Note that no other assumptions are made about either the distribution of the $\varepsilon_{d,h,m}$ or their serial dependence.[2] The $\alpha_h$ parameters can be estimated easily using ordinary least squares to fit this zero-intercept multiple linear regression model to increment data. The variance parameters $\sigma_m^2$ can then each be estimated by the sample variances of the $\hat{\varepsilon}_{d,h,m}$ residuals from the model of best fit. By fitting this model, it is then possible to determine how much of the MAE of

[2] For $\varepsilon_{d,h,m}$ that are well represented by an autoregressive 1 (AR1) process, it can be shown that $\sigma_m^2 = \sigma^2(1 - \rho^{2m})/(1 - \rho^2)$ for $m > 0$, where $\rho$ is the lag 1-min autocorrelation and $\sigma^2$ is the variance of the AR(1) residuals. This was found to agree well with the sample estimates shown in the following section for the temperature data—the decorrelation times—$(\log \rho)^{-1}$ were found to be around 20–25 min.

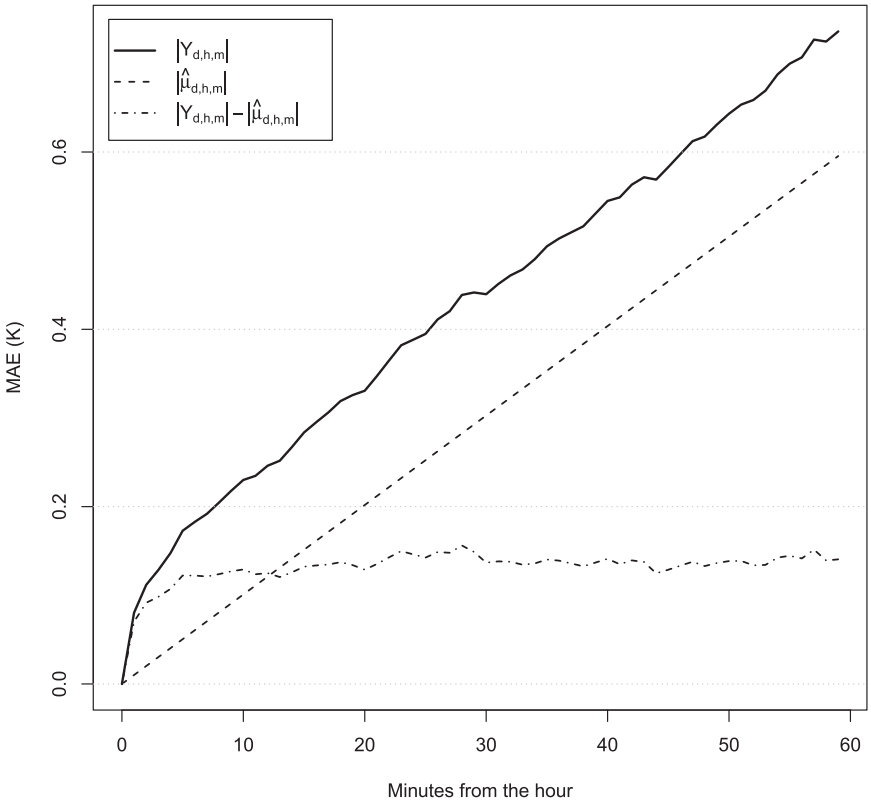


FIG. 4. Increase in the total temporal uncertainty as a function of minutes from the hour, and the contribution of the trend $|\hat{\mu}_{d,h,m}|$ and the irreducible part $|Y_{d,h,m}| - |\hat{\mu}_{d,h,m}|$ for Heathrow, June 2013. In this case $\overline{|Y_{d,h,m}| - |\hat{\mu}_{d,h,m}|} = 0.13$ K.

perfect forecasts is due to trend caused by the diurnal cycle [i.e., the sample mean of $|\hat{\mu}_{d,h,m}|$ (a diurnally varying bias that in principle can be removed by postprocessing)], and how much is due to irreducible noisy variations within an hour (i.e., the sample mean of $|Y_{d,h,m}| - |\hat{\mu}_{d,h,m}|$). The following section will test the validity of this modeling approach and will then use it to diagnose the impact of subhourly variations on forecast accuracy.

## 4. Results: U.K. temperature example

For this study the temperatures are dithered with a small amount ($\pm$ 0.05) of uniformly distributed noise to mitigate for the observations being recorded only to the nearest 0.1 K. Throughout most of this section the focus is on the 1-min temperature values for Heathrow during June 2013. Figure 3 contains a range of diagnostics to illustrate the goodness of fit of the multiple linear regression model outlined in the previous section. Figure 3a demonstrates that the residual increments $Y_{d,h,m}$ are independent of the day of the month, while Fig. 3b illustrates the behavior as a function of time of day. The variance $\sigma_m^2$ as a function of minutes from the hour is shown in Fig. 3c. The line indicates $\pm$ 3 standard deviations, which for a normal distribution should overlap 99.7% of the values. The near-normal behavior of residuals is evident from Fig. 3d, which shows the density function for a 30-min offset.

The total temporal uncertainty $|Y_{d,h,m}|$ as a function of minutes from the top of the hour is shown in Fig. 4. The linear trend $|\hat{\mu}_{d,h,m}|$ is a substantial component. As stated earlier, this part is in principle reducible, using appropriate postprocessing methods such as the KF method. However, a mean irreducible part of 0.13 K remains. Even perfect forecasts are exposed to the full impact of $|Y_{d,h,m}|$, in this case 0.22 and 0.44 K for a 10- and 30-min offset, respectively.

The model described in section 3 was fitted to minute-resolution time series at the seven selected sites for each month between August 2012 and June 2013. Figure 5 shows that the $m = 30$ standard deviations $\sigma_\epsilon$ depends on location and time of year. Southwest-facing coastal locations such as St. Mary's and St. Athan show little annual variation, while other locations exhibit a distinct annual cycle with the largest values in the warmer seasons, and a minimum in January.

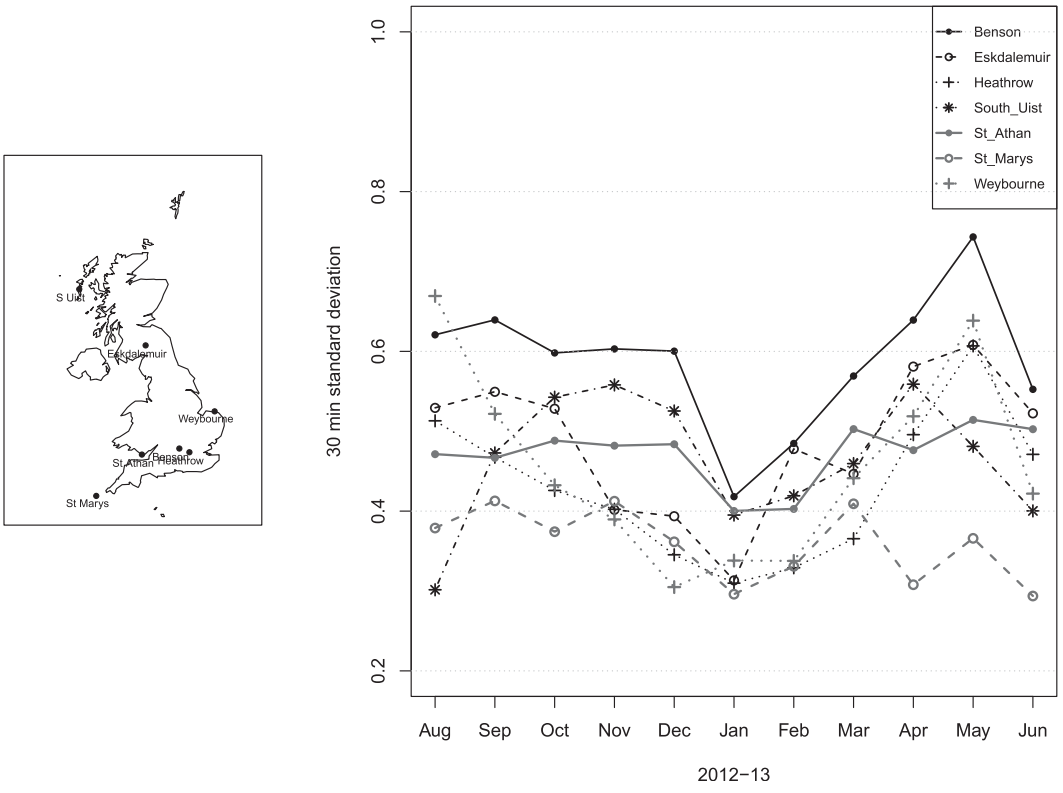


FIG. 5. (right) Monthly standard deviations of $|Y_{d,h,m}|$ (in K) for a 30-min offset for the seven locations studied, with (left) a map to indicate the location of the sites.

## 5. Conclusions

A flexible yet robust method for quantifying the observation uncertainty associated with temporal sampling is demonstrated by applying it to 1-min temperature time series (with the understanding that currently a maximum temporal offset of 10 min exists between synoptic observations and model output). The results show there is an irreducible uncertainty component that is at least of the same order of magnitude as the instrument measurement error for surface temperature, imposing a nonzero lower bound on the achievable level of temperature forecast accuracy. Compared to Bowler (2006) the method shows that using the hourly observation as a "perfect" forecast, an MAE of zero is possible only if there is no temporal mismatch between the observation and the forecast.

Further work will focus on other variables and the impact of probabilistic scores. It may also be worth investigating how this uncertainty also affects data assimilation and initialization; for example, would it be better if the observations were assimilated at the exact time they were taken? Finally, this source of uncertainty depends on the serial dependency in the weather variables and so is likely to change under different climatic conditions (e.g., more persistent temperatures during prolonged droughts).

*Acknowledgments.* Marion Mittermaier would like to thank NCAR-DTC for their visiting scientist grant support used to explore this subject. Further collaboration was kindly enabled by the Met Office Academic Partnership secondment scheme (http://www.metoffice.gov.uk/research/partnership).

## REFERENCES

Bowler, N., 2006: Explicitly accounting for observation error in categorical verification of forecasts. *Mon. Wea. Rev.*, **134**, 1600–1606, doi:10.1175/MWR3138.1.

——, 2008: Accounting for the effect of observation errors on verification of MOGREPS. *Meteor. Appl.*, **15**, 199–205, doi:10.1002/met.64.

Candille, G., and O. Talagrand, 2005: Evaluation of probabilistic prediction systems for a scalar variable. *Quart. J. Roy. Meteor. Soc.*, **131**, 2131–2150, doi:10.1256/qj.04.71.

——, and ——, 2008: Impact of observational error on the validation of ensemble prediction systems. *Quart. J. Roy. Meteor. Soc.*, **134**, 959–971, doi:10.1002/qj.268.

Hansen, J., R. Ruedy, M. Sato, and K. Lo, 2010: Global surface temperature change. *Rev. Geophys.*, **48**, RG4004, doi:10.1029/2010RG000345.

Jolliffe, I. T., and D. B. Stephenson, Eds., 2011: *Forecast Verification: A Practitioner's Guide in Atmospheric Science*. 2nd ed. John Wiley and Sons, 292 pp.

Koh, T.-Y., B. Bhatt, K. Cheung, C. Teo, Y. Lee, and M. Roth, 2012: Using the spectral scaling exponent for validation of

quantitative precipitation forecasts. *Meteor. Atmos. Phys.*, **115**, 35–45, doi:10.1007/s00703-011-0166-4.

Morice, C., J. J. Kennedy, N. A. Rayner, and P. D. Jones, 2012: Quantifying uncertainties in global and regional temperature change using an ensemble of observational estimates: The HadCRUT4 data set. *J. Geophys. Res.*, **117**, D08101, doi:10.1029/2011JD017187.

Röpnack, A., A. Hense, C. Gebhardt, and D. Majewski, 2013: Bayesian model verification of NWP ensemble forecasts. *Mon. Wea. Rev.*, **141**, 375–387, doi:10.1175/MWR-D-11-00350.1.

Saetra, O., H. Hersbach, J.-R. Bidlot, and D. Richardson, 2004: Effects of observation errors on the statistics for ensemble spread and reliability. *Mon. Wea. Rev.*, **132**, 1487–1501, doi:10.1175/1520-0493(2004)132<1487:EOOEOT>2.0.CO;2.

Santos, C., and A. Ghelli, 2012: Observational uncertainty method to assess ensemble precipitation forecasts. *Quart. J. Roy. Meteor. Soc.*, **138**, 209–221, doi:10.1002/qj.895.

Smith, T., R. Reynolds, T. Peterson, and J. Lawrimore, 2008: Improvements to NOAA's historical merged land–ocean surface temperature analysis (1880–2006). *J. Climate*, **21**, 2283–2296, doi:10.1175/2007JCLI2100.1.

WMO, 2008: Guide to meteorological instruments and methods of observation. Tech. Rep. WMO-8, World Meteorological Organization, Geneva, Switzerland, 713 pp. [Available online at http://library.wmo.int/pmb_ged/wmo_8_en-2012.pdf.]

——, 2014: Manual on codes—International codes. Vol. I.1: Part A—Alphanumeric codes. Tech. Rep. 306, World Meteorological Organization, Geneva, Switzerland, 466 pp. [Available online at: https://drive.google.com/file/d/0BwdvoC9AeWjUOFdkaXlCUUpETmM/view?usp=sharing.]