Quarterly Journal of the Royal Meteorological Society Q. J. R. Meteorol. Soc. 141: 2923–2933, October 2015 B DOI:10.1002/qj.2612

RMetS
Royal Meteorological Society

Featured Article

Decoding the precision of historical temperature observations

Andrew Rhines,[a]* Martin P. Tingley,[b] Karen A. McKinnon[a] and Peter Huybers[a]
[a]*Department of Earth and Planetary Sciences, Harvard University, Cambridge, MA, USA*
[b]*Departments of Statistics and Meteorology, Pennsylvania State University, State College, USA*

*Correspondence to: A. Rhines, Department of Earth and Planetary Sciences, Harvard University, 20 Oxford Street, Cambridge, MA 02138, USA. E-mail: arhines@fas.harvard.edu

Historical observations of temperature underpin our ability to monitor Earth's climate. We identify a pervasive issue in archived observations from surface stations, wherein the use of varying conventions for units and precision has led to distorted distributions of the data. Apart from the original precision being generally unknown, the majority of archived temperature data are found to be misaligned with the original measurements because of rounding on a Fahrenheit scale, conversion to Celsius, and re-rounding. Furthermore, we show that commonly used statistical methods including quantile regression are sensitive to the finite precision and to double-rounding of the data after unit conversion. To remedy these issues, we present a Hidden Markov Model that uses the differing frequencies of specific recorded values to recover the most likely original precision and units associated with each observation. This precision-decoding algorithm is used to infer the precision of the 644 million daily surface temperature observations in the Global Historical Climate Network database, providing more accurate values for the 63% of samples found to have been biased by double-rounding. The average absolute bias correction across the dataset is 0.018 °C, and the average inferred precision is 0.41 °C, even though data are archived at 0.1 °C precision. These results permit better inference of when record temperatures occurred, correction of rounding effects, and identification of inhomogeneities in surface temperature time series, amongst other applications. The precision-decoding algorithm is generally applicable to rounded observations–including surface pressure, humidity, precipitation, and other temperature data–thereby offering the potential to improve quality-control procedures for many datasets.

Key Words: quality control; surface stations; instrumentation; metadata; precision

Received 19 March 2015; Revised 9 June 2015; Accepted 23 June 2015; Published online in Wiley Online Library 08 September 2015

1. Introduction

Daily temperature maxima and minima measured at surface weather stations are the primary observational source for information about the last two centuries of climate, forming the bulk of pre-satellite observations (Thorne *et al.*, 2011). Historical measurements of daily temperature extrema are more reliable (Wang, 2014) and widely available than daily averages because self-registering thermometers–which have been in widespread use since the 1800s–can track the diurnal range of near-surface temperature without continuous human observation (Austin and McConnell, 1980). Archival, restoration, and digitization of these time series has been crucially important for studying historical weather and climate variations (e.g. Brohan *et al.*, 2006).

The Global Historical Climate Network Daily (GHCND) database (Menne *et al.*, 2012) is the largest aggregation of quality-controlled daily surface observations. A number of biases and errors in these time series are known to stem from physical changes in the observing system, such as by station relocation (Feng *et al.*, 2004), from human factors that cause certain numbers to be preferentially chosen under uncertainty (Nese, 1994), and from transcription or digitization errors that occur during the archival process (Reek *et al.*, 1992; Torok and Nicholls, 1996; Durre *et al.*, 2010). Existing quality-control procedures have prevented many erroneous samples from being included in analyses, though these have primarily focused on identifying implausible values such as long strings of zeros, or unrealistic values relative to those of nearby stations (Durre *et al.*, 2010; Muller *et al.*, 2013), with many techniques being applied only to monthly averages.

Observational precision is an important source of uncertainty that has not been adequately addressed. Although all GHCND observations are archived in increments of 0.1 °C, their original precision and units are not generally reported, and may vary in time and from station to station. Many individual time series are well-documented, but metadata from other stations are absent, incomplete, or non-standardized (e.g. as handwritten

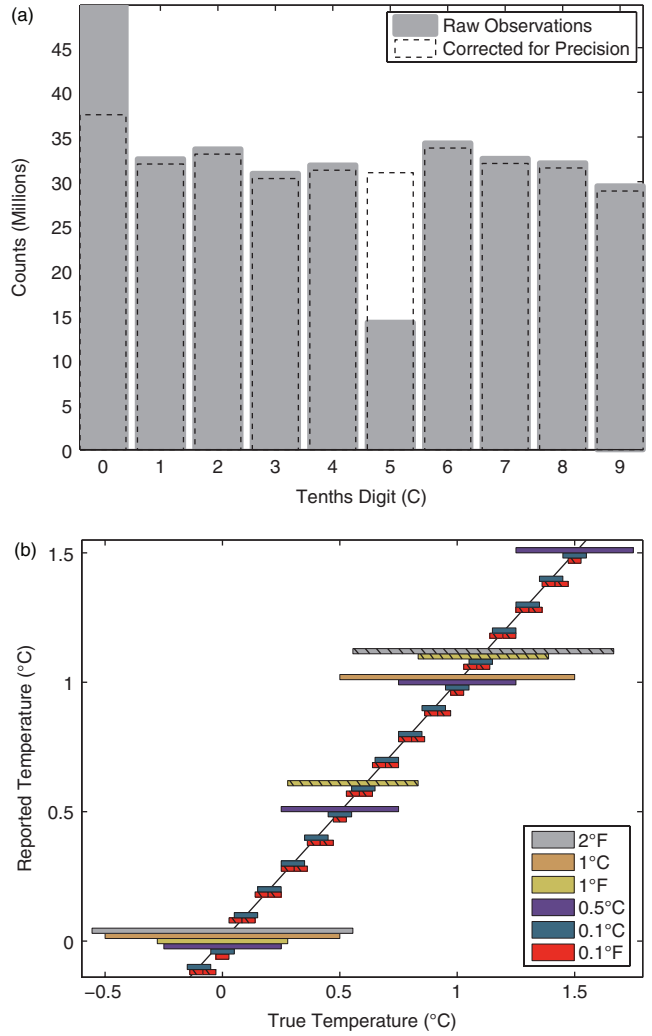

Figure 1. (a) The effect of unit conversion and double-rounding errors on daily maximum temperature observations from the GHCND database. The tenths digit is not uniformly distributed (solid) as a result of variable precision and units in the original data. Correcting for the presence of different precisions and units detected by precision decoding, the distribution is found to be nearly uniform (dashed), though human-induced biases due to preferential recording of the zero digit still appear to be an issue. Panel (b) illustrates how rounding and unit conversion alters the archived temperature for various original precisions. Hatching indicates that the archived value is offset from the mean true temperature on that interval, whereas unhatched bars indicate that the interval is correctly centered. Note that each coloured sequence forms a non-overlapping staircase covering all true temperatures. Bars are slightly offset in the vertical for visibility.

notes), and the diversity of data sources makes their recovery difficult for large, global datasets (Durre *et al.*, 2010). In some cases it is possible to locate the original recording medium–as with older data from Cooperative Observer Program stations, derived from forms that instruct the operator to record values in Fahrenheit to the nearest integer. Although many raw digitized observations have been made available through the heterogeneous source data compiled by the Global Land Surface Databank (GLSD; Thorne *et al.*, 2011), in certain cases there is no readily identifiable information regarding precision. For example, the lowest-level form of data for a station in Mombasa, Kenya (GHCND station code KE000063820) in the GLSD is a set of manually aggregated monthly average temperature tables with little supporting information beyond the station location and a list of numbers and dates. The units can be inferred through comparison with the regional climatology, but the precision must be inferred in some other way.

In addition to being an uncharacterised source of uncertainty, data that have been double-rounded–that is, rounded in the initial measurement units, possibly converted to different units, and then rounded a second time–can lead to a surprising variety of errors in subsequent analyses (Figueroa, 1995). Evidence of a substantial double-rounding effect can be seen in the frequency of each possible tenths digit in the GHCND data (Figure 1(a)), in which zeros are systematically overrepresented (comprising 14% of all samples) and fives are under-represented (comprising only 4% of samples). Unlike the leading digit, which is logarithmically distributed in most data (Durtschi *et al.*, 2004), the trailing digit is typically uniformly distributed (Preece, 1981), with deviations stemming from distorted sampling (Al-Marzouki, 2005). The inflated zero counts alone might be explained by digit preferences of station operators or by some observations having originally been 1.0 °C-precision, but the under-representation of fives suggests a further effect given the large sample size and the relatively uniform frequency of the remaining digits.

Observations originally recorded to a precision of 1.0 °F, converted to Celsius and then rounded again to 0.1 °C-precision never feature a five in the decimal position (Figure 1(b)). For example, integer values between 32 and 42 °F convert to 0.1 °C-precision values of 0.0, 0.6, 1.1, 1.7, 2.2, 2.8, 3.3, 3.9, 4.4, 5.0, and 5.6 °C, respectively, in a repeating sequence of decimal digits. The under-representation of fives (Figure 1(a)) can therefore be understood as a consequence of the GHCND database containing some data that were originally of 1.0 °F-precision, a fact that has been previously noted (Zhang *et al.*, 2009), but for which the consequences of double-rounding have not yet been assessed. Furthermore, there is no established method to automatically infer the original precision even in the relatively simple case of a constant rounding and conversion protocol, though of course this can also vary in time.

In the following, we introduce an algorithm termed precision decoding which accurately infers the original precision and units of observations (section 2), and then demonstrate its operation upon synthetic data (section 3) and present results and implications (section 4) as well as conclusions (section 5).

2. Precision decoding

The objective of the precision-decoding algorithm is to determine the most likely precision and units of each sample in a given time series. The algorithm works by determining whether a given sequence of observations is consistent with each of several candidate precision levels, modelling the unobserved precision and units as a latent variable which is then inferred using a Hidden Markov Model (HMM). The algorithm makes use of the fact that each method of observing, rounding, and converting physical observations leaves distinct fingerprints in double-rounded data in the form of predictable distributions of possible values (Figures 1 and 2). Precision decoding recovers time series of the most likely precision and units of the original observation and, in cases where double-rounding has misaligned the archived values, recovers the original observations and the range of possible values implied by the inferred precision, also known as the preimage.

Although the algorithm is general and can be applied to any time series, for purposes of specificity, we describe it with reference to temperature data from the Global Historical Climate Network Daily database (GHCND; Menne *et al.*, 2012). The GHCND data are distributed as integer multiples of 0.1 °C, in accord with the format recommended by the World Meteorological Organization. Each measurement is provided with one or more flags denoting metadata or quality-control information, and we only consider minimum and maximum temperature values without negative quality-control flags.

Precision is generally not directly observable, but rather must be detected from the data, motivating the use of an HMM to estimate the most likely precision state through time. A solution for this Markov model which is optimal in the maximum *a posteriori* sense is obtained through the Viterbi algorithm (Forney, 1973). The algorithm requires the specification of an emission matrix linking the precision state to the observations, **E**, a transition matrix describing the likelihood of the precision or

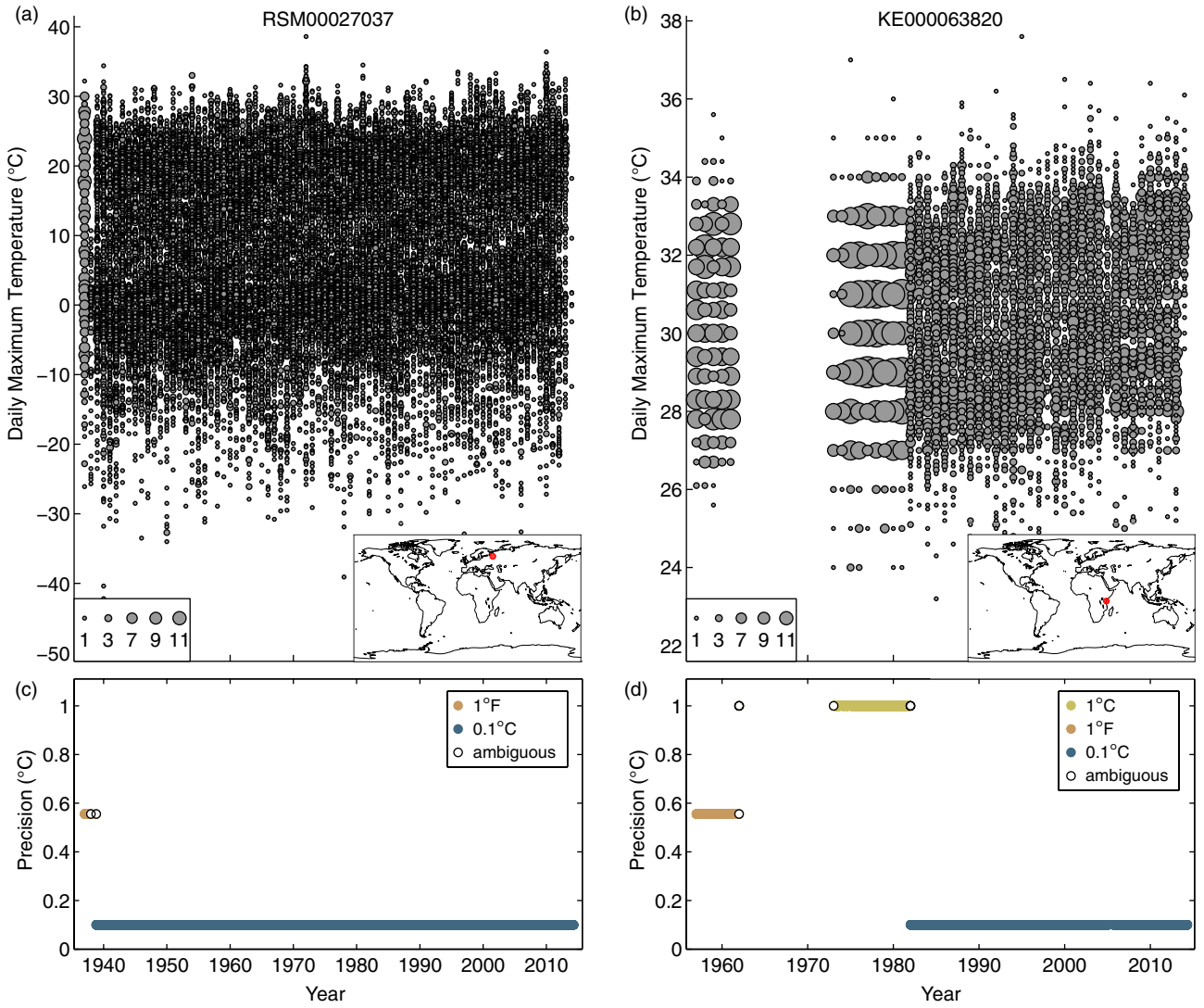

Figure 2. Detection of historical precision in two example surface stations. Daily maximum temperature data from stations in (a) Russia (59.32°N, 39.92°E) and (b) Kenya (4.03°S, 39.62°E) are plotted by year with circle area proportional to the number of observations recorded at a given value, illustrating that the data lie on quantized levels that vary over time. (c,d) Precision changes detected using precision decoding show that early observations in (a) were taken at 1 °F precision, followed by more accurate 0.1 °C observations. In (b), the record also begins with 1 °F precision, followed by a hiatus and a transition to 1 °C when observations resumed in 1973, finally switching to 0.1 °C in 1982. Points with ambiguous precision within the likely range of a detected state transition are indicated in white.

units changing, **A**, an initial state distribution, Π, and candidate combinations of precision, X. The convention used here is that E_{lk} gives the likelihood of emission l when the system is in state k, and A_{ij} gives the probability of transitioning from state i to state j (Figure 3). Π_i represent the probability that the system is initially in state i. We assume that the system satisfies the Markov property, i.e. that the system's current state and emissions depend only on its most recent state. If the transition, emission, and initial state probabilities are known exactly and the system satisfies the Markov property, the Viterbi algorithm determines the most likely state sequence, optimal in the sense of maximum *a posteriori* probability. In the more general case that we are confronted with model parameters that are unknown, they can be estimated iteratively using the Baum–Welch algorithm (Rabiner, 1989). As the number of precision states present in a specific time series is not known *a priori*, we also use the Bayesian Information Criterion (BIC; Schwarz, 1978) to identify the optimal model and avoid overfitting by using too many states. Using the parameters obtained from Baum–Welch augmented by the BIC, the Viterbi algorithm then provides an estimate of the most likely state sequence. These computational methods are widely used in signal processing, machine learning, bioinformatics, and other areas of research (e.g. Durbin *et al.*, 1998).

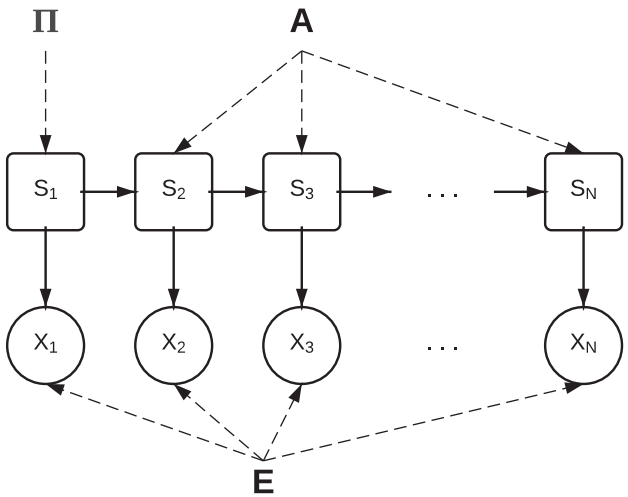

Figure 3. State-space representation of the HMM used in precision-decoding. The unobserved precision state sequence, S, is distributed according to the initial state probabilities, Π, with transitions governed by the transition matrix, **A**. At each time $j = \{1, 2, \ldots, N\}$, the categorical variable X_j represents the mapping from the archived temperature to the precision consistency. The likelihood of each precision level being present, conditional on the unobserved state, is given by the emission probability matrix, **E**.

2.1. Precision candidates and consistency subsets

Use of an HMM requires a mapping from the temperature time series to a time series representing consistency with each candidate precision state, as the Markov property is not satisfied unless the emission probabilities depend only on the current precision state. Raw temperature values are generally non-Markovian in that they follow a seasonal cycle and thus explore only a small range of possible emissions at any given time, hence the need for a transformation. We make use of the fact that each assumed original precision level implies a distinct distribution of archived 0.1 °C-precision values, and, conversely, each archived value at 0.1 °C-precision is only possible under a subset of the original candidate precision levels (Figure 1(b)). By identifying the original precision level and assuming conventional rounding and unit conversions, we can identify the range of values in the original precision that correspond to the archived GHCND value.

We assess the consistency of each observation, T_j, with candidate precisions indexed by k. Consistency is tabulated in a matrix, **C**, denoting whether T_j could have been recorded using precision state k. Specifically, for Celsius precision states having a precision of p_k decimal digits, the consistency matrix is defined as

$$c_{jk} = \begin{cases} d & \text{if } T_j \equiv 0 \pmod{10^{p_k}}, \\ 0 & \text{otherwise,} \end{cases} \tag{1}$$

where d is the degeneracy–the number of unique values in the original precision that map onto T_j. For example, d is ten if the original precision is hundredths of a degree Celsius and the final precision is tenths of a degree Celsius. Cases where data are archived in Celsius but originally recorded in Fahrenheit require a modified inversion procedure which is described in section 2.2. Many rows of **C** are identical, and we refer to each unique row as a precision consistency subset. Each of these subsets acts as a symbol that carries information about the underlying precision state, and the rate at which each symbol occurs is the basis for decoding the precision using an HMM.

With R candidate precisions there are up to $2^R - 1$ possible consistency subsets, but the number is much smaller in practice due to repeating patterns which eliminate many combinations. For example, every temperature consistent with 1.0 °C-precision is also consistent with 0.5 and 0.1 °C, while the converse is not true. The use of consistency subsets implicitly assumes that day-to-day temperature variability is large compared with typical precision increments, in that the symbol emitted at time t_j is assumed to be independent of that emitted at time t_{j-1}, conditional on the true precision used for each sample. If the data are strongly autocorrelated such that successive values are likely to be identical, the assumption of independence could be relaxed by considering only every τth sample, where τ is an empirical decorrelation increment. In practice, we find this additional step to be unnecessary. (For example, the Majuro Airport station in the Marshall Islands (GHCND code RMW00040710, discussed in Zhang *et al.*, 2009) has strong autocorrelation and reported only 17 unique values of both minimum and maximum temperature, with identical temperatures being reported for consecutive days approximately 25% of the time. While the distribution of decimal digits is distorted compared with the expected distribution for high-variability stations, in the space of consistency subsets the Fahrenheit data remain clearly distinguishable from other precisions until 2013 after which the station apparently reports at 0.1 °C precision.

We use candidate precisions of 0.1, 0.5, and 1.0 °C, as well as 0.1, 1.0, and 2.0 °F. The 0.1 °C, 1.0 °C, and 1.0 °F candidates are chosen because they are established standards documented to have been used by different station networks. The additional possibilities of 0.5 °C, 0.1 °F, and 2.0 °F were included following exploratory analysis which suggested additional precisions were necessary to explain the results at a large number of stations. We do not exclude that there may be rare cases with additional precisions, or that the effective precision may be lower where digit-preference biases are present (Nese, 1994). This set of candidates yields a relatively small set of seven possible subsets, while maintaining a high degree of distinguishability between different candidate precisions. Having a small number of candidate states as well as a small number of subsets is also advantageous from the perspective of computation, as the Viterbi and Baum–Welch algorithms have complexity $\mathcal{O}(N \times R^2 + N \times R \times K)$, where N is the number of samples in the time series, R is the number of precision candidates, and K is the number of possible consistency subsets.

Missing values are ignored here rather than being assigned a special consistency category, a simplification that has some immediate–but minor–effects. As a result of ignoring the gap length, the ambiguity of transition points can spread across the observation gaps (e.g. Figure 1(d)). Simple cases of this type can be corrected for in post-processing, and extensions to the algorithm could also be made to explicitly model gaps. Because gaps in observations may coincide with observing system changes, it would ultimately be useful to address the missing values in the state estimates; the Viterbi algorithm does not permit for missing observations, though alternative methods such as posterior decoding (Durbin *et al.*, 1998) could be used to augment the Viterbi results in such cases.

2.2. Inversion given both rounding and unit conversion

Computation of the consistency between observations and candidate precision levels requires the identification, for each candidate precision, of the range of values that map onto each value archived in the GHCND database. This mapping can be derived by first considering the rounding function,

$$\text{round}(x) = \left\lfloor x + \frac{1}{2} \right\rfloor, \tag{2}$$

where the floor function ($\lfloor \cdot \rfloor$) is defined by the inequality for some integer n,

$$\lfloor x \rfloor = n \Leftrightarrow n \leq x < n + 1. \tag{3}$$

Note that the convention to round upwards used here is consistent with meteorological conventions but differs from alternative definitions sometimes used in other fields, for example rounding away from 0 at half-integer values, or rounding to the nearest even integer (so-called *bankers' rounding*, designed so that negation commutes with rounding). We also note that, for the temperature units considered here, all rounding methods give identical results with precision-decoding.

For a destination precision of p decimal digits, the true value of the observed temperature, T, is rounded to

$$\left\lfloor 10^p T + \frac{1}{2} \right\rfloor 10^{-p}. \tag{4}$$

In general, the original temperature may have been rounded multiple times and is perhaps subject to unit changes as well. Any such sequence of rounding and data processing steps can be represented as a composition of similar functions.

The rounding function is many-to-one. Although the function is not invertible, it is possible to solve for its preimage (the range of possible values prior to rounding) if the data processing sequence is known. As a concrete example, consider the question of whether a given archived Celsius observation, T_r, with p_r decimal digits of precision, could derive from an original Fahrenheit observation with p_o digits of precision. Repeated application of Eqs (3) and (4) and the conversion from Fahrenheit to Celsius lead to the inequality,

$$\begin{aligned} \frac{9}{5}\left(T_\text{r} - \frac{10^{-p_\text{r}}}{2}\right) + 32 \;&\leq n \times 10^{-p_o} \\ &< \frac{9}{5}\left(T_\text{r} + \frac{10^{-p_\text{r}}}{2}\right) + 32, \end{aligned} \tag{5}$$

where $n \in \mathbb{Z}$ and $n \times 10^{-p_o}$ is the originally recorded temperature in Fahrenheit. For a given p_o and p_r, all n satisfying this inequality must be found to evaluate consistency. A solution exists only if at least one integer n satisfies the inequality, a fact that provides a strong constraint for inferring precision. Multiple solutions can exist wherever the interval has a width greater than 10^{-p_o}. The number of solutions (the degeneracy in Eq. (1)), d, is then bounded by

$$\left\lfloor \frac{9}{5} 10^{p_o - p_r} \right\rfloor \leq d \leq \left\lceil \frac{9}{5} 10^{p_o - p_r} \right\rceil, \tag{6}$$

where $\lceil \cdot \rceil$ is the ceiling function. For example, if the Celsius precision is one decimal digit (p_r equals one, as in the GHCND database), and the Fahrenheit precision is 1 °F (p_o equals 0), there is at most one solution and the inverse can be recovered in every case where the conversion was from a precision of 1 °F to 0.1 °C. Often, however, double-rounding leads to irreversible errors (Figueroa, 1995), as a result of degeneracy if the original precision is smaller than the archive precision−80% of all 0.1 °C values have two possible 0.1 °F source values, hence the need for a treatment of degeneracy in the consistency subsets.

Equation (5) can be solved explicitly using integer linear programming, but in practice a more efficient solution is to solve the forward problem on an exhaustive list of values in the destination units, and then check for consistency via a look-up table which we implement as a binary search tree.

2.3. Emission matrix

The emission matrix, **E**, has entries, E_{lk}, denoting the likelihood of observing consistency subset l when the true observational precision state index is k. For example, if the only possible precisions are $r_1 = 0.1$ °C and $r_2 = 1$ °C, the only consistency subsets possible for all observations are $l_1 = \{0.1\,°C\}$ and $l_2 = \{0.1, 1\,°C\}$ because there are no values that could be consistent with the precision being 1 °C without also being consistent with 0.1 °C. In this truncated case, the full emission likelihood is

$$\mathbf{E} = \begin{pmatrix} 9/10 & 0 \\ 1/10 & 1 \end{pmatrix}. \tag{7}$$

For convenience and interpretability, the emission likelihood is normalized down columns of **E**, so that each entry represents the probability of emitting subset l from precision state k. In computing **E**, we use a flat prior for the true temperature, i.e. we assume that all temperature values are equally likely. This is equivalent to the assumption that temperature variability is large relative to the precision increments. Real temperature time series explore a finite range ($\mathcal{O}(100\,°C)$), so that the asymptotic limit is only approximate. Nevertheless, the emission likelihoods are sufficiently distinct that we do not find any improvement in classification in synthetic tests when a smoothed empirical distribution specific to each time series is used to generate **E** instead of the flat prior; the likelihoods are typically higher, but the decoded state sequence is indistinguishable as a result of the consistency mapping providing a strong constraint.

2.4. Parameter estimation

We use the Baum−Welch algorithm (Rabiner, 1989) to provide an estimated transition matrix, **A**, and initial state probabilities, Π. The Baum−Welch algorithm uses expectation maximization (Dempster *et al.*, 1977) to converge on a maximum likelihood estimate of the model parameters given the observations.The probability of transitioning from one observational precision to another is generally unknown and can be expected to vary substantially from station to station, but is assumed to be relatively unlikely. To reflect this, we initialize the algorithm with

$$\hat{\mathbf{A}} = \begin{cases} \alpha & \text{if i=j,} \\ \dfrac{1-\alpha}{R-1} & \text{otherwise.} \end{cases} \tag{8}$$

We set α equal to 0.999 (the Baum−Welch algorithm is not sensitive to this choice, but converges more quickly when initialized with large self-transition probabilities). We add uniform, randomly generated noise on the interval $(0, 10^{-5})$ to each entry and re-normalize the matrix to avoid issues with initial symmetry known to influence the outcome of the subsequent parameter estimation (Durbin *et al.*, 1998). Because we have an analytical solution for the emission probabilities, we have modified the standard Baum−Welch implementation so that it estimates only the transition matrix and the initial state probabilities. We use a tolerance of 10^{-6} and at most 500 iterations for convergence.

Maximum likelihood estimates conditional on the parameter estimates are obtained for each possible combination of precision candidates. The optimal set of precision candidates is selected as the one that minimizes the BIC (Schwarz, 1978),

$$\mathrm{BIC} = -2L + D\log(N), \tag{9}$$

where L is the log-likelihood of the data given the model with maximum-likelihood parameters as estimated by the Baum−Welch algorithm, D is the number of independent parameters estimated, and N is the number of observations in the time series. D depends only on the number of precision states assumed to be present, R:

$$D = (R-1) + R(R-1). \tag{10}$$

Because we do not train the emission probabilities, the degrees of freedom stem from the transition matrix, **A**, contributing $R(R-1)$ since each row must sum to one, and the initial state probability, Π, contributing $R-1$ since Π must also sum to one.

In most cases the standard form of the criterion is adequate, however we note that in cases where there are biases in recording practices, some further *ad hoc* measures are required to prevent over-fitting. Specifically, when decimal digits of 5 and 0 are preferentially recorded, the parameter estimation step may compensate by jumping into and out of the 0.5 °C state, e.g. from the 0.1 °C state. To prevent such flickering, we make a modification if the model chosen by the Bayesian Information Criterion predicts more than ten state transitions, instead using a modified Bayesian Information Criterion (mBIC),

$$mBIC = -2L + (D+W)\log(N), \tag{11}$$

where W is the number of transitions in the Viterbi path. The mBIC is similar to the standard BIC for change point methods in which the model complexity scales with the number of transitions rather than the number of states (Bogdan *et al.*, 2008). We find that the results are not overly sensitive to the threshold choice of ten transitions, as the mBIC is primarily invoked where severe overfitting results in transitions occurring regularly. However, in order to avoid missing potentially important cases where a small but non-zero number of transitions are present (as in Figure 2), we use the mBIC only above this threshold. The modified criterion is used for 1947 of the 28 824 examined stations. These stations are geographically widespread, with the highest concentrations in the USA due to the overall high station density there. Another possible contributing factor is the prevalence in the USA of many longer time series, and therefore presumably a greater susceptibility to human digit preferences in the older observations. If hard classification decisions are not required, multiple selection criteria including the BIC and mBIC could be used simultaneously in

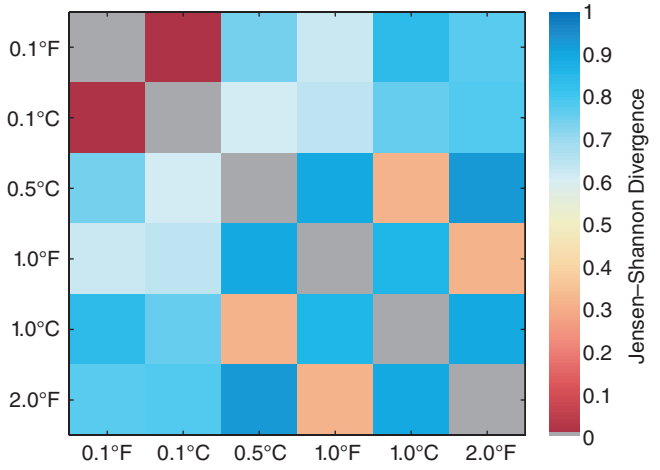

Figure 4. The Jensen–Shannon divergence between different precisions in consistency space. Small values indicate that a pair of states is difficult to distinguish, whereas large values indicate that the states can be easily distinguished with a small number of samples. Values exactly equal to 0 are in grey.

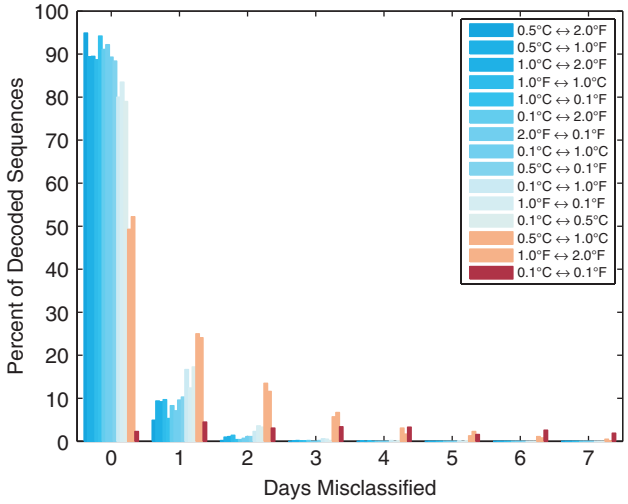

Figure 5. Histograms of the error rate in the decoded precision of synthetic observations. For each pair of precisions, 1000 trials using 20 years of synthetic data are generated, with the observational precision changing after 10 years. Histogram bars are shaded by the Jensen–Shannon divergence on the same colour scale as in Figure 4. Non-zero errors in the 0.1 °F-to-0.1 °C case occur over a much wider range which is not fully shown, with an expected misclassification count of 35 days. Its distribution has a qualitatively similar shape to the other cases.

a hierarchical setting to provide a smoothed state estimate that represents the contribution of several models. Based on results from synthetic tests (following section), we also forbid direct transitions between the 0.1 °F state and either 0.1, 0.5, or 1 °C. As with the mBIC, this choice reduces overfitting and only appears to influence a small number of stations that have large digit-preference biases.

3. Validation of the algorithm

We perform several synthetic tests of the algorithm to evaluate its performance on realistic sample sequences. The tests are performed using both unbiased and biased samples to determine the robustness of the algorithm. The resolution of the model selection procedure is also tested by varying the length of the analysed sequences.

In evaluating the algorithm's performance it is important to account for the fact that some pairs of precisions are more difficult to distinguish than others; some pairs often produce mutually consistent samples (e.g. 0.5 and 1.0 °C), while others do so less frequently (e.g. 0.5 °C and 1.0 °F). It is therefore useful to define a distinguishability measure to evaluate the algorithm relative to an information-theoretic baseline. A symmetric measure of the difference between two discrete distributions is the Jensen–Shannon Divergence (JSD; Lin, 1991),

$$JSD(Q||P) = \frac{1}{2}\sum_i Q_i \log\left(\frac{2Q_i}{Q_i+P_i}\right) + \frac{1}{2}\sum_i P_i \log\left(\frac{2P_i}{Q_i+P_i}\right), \quad (12)$$

where P and Q are two discrete distributions–here, in the space of consistency subsets. By construction the JSD values fall between 0 and 1, and are an upper bound for the rate at which information in the observations can be used to distinguish between two possible states. Small values (e.g. 1.1×10^{-3} for the {0.1 °C, 0.1 °F} comparison) indicate that the two states in question are difficult to distinguish given a set of observations, and large values (e.g. 0.93 for the {0.5 °C, 2.0 °F} comparison) indicate that states are easier to distinguish. Most candidate precision pairs are readily distinguishable (Figure 4), though differentiating 0.1 °C from 0.1 °F is considerably more difficult. We note that at archival precisions smaller than 0.1 °C, the 0.1 °F and 0.1 °C states would be more clearly distinguishable. In the tests that follow, we stratify the algorithm's performance according to the JSD of the two precisions being compared.

Synthetic data are generated to resemble autocorrelated daily temperature observations with a seasonal cycle. The time series are generated as an AR(1) process with noise variance of 9 °C^2 and an autoregressive coefficient of 0.8, which is then added to a 20 °C amplitude sinusoidal seasonal cycle, representative of a typical midlatitude location. Variation of the AR(1) parameters and the seasonal cycle amplitude over a wide range representative of different climatic zones suggests that the results are not sensitive to these choices. We sample the data using different sequences of precision and unit combinations, and evaluate the accuracy of precision-decoding in identifying the original precision of each synthetic time series.

First, we generate 20-year-long sequences with two precisions present, each comprising half of the sequence so that the transition occurs in the middle of the sequence. As is common with HMM state inference problems (Durbin *et al.*, 1998), errors tend to occur in 'bursts' so that the misclassification distribution is positively-skewed. Errors for each case appear to approximately follow a Poisson distribution whose parameter is related to the case's Jensen–Shannon Divergence. When the sampling is unbiased, the misclassification count has a median of 0 and a mean of between 0 and 1 days for most cases, albeit with some notable exceptions. The combination of 0.1 °F and 0.1 °C results in an average misclassification of 35 days, and approximately 1.5 days for the 0.5 °C versus 1.0 °C and 1.0 °F versus 2.0 °F pairs (Figure 5).

Robustness of the algorithm to digit preference biases is assessed by imposing a random preference in the original units towards trailing digits of {5, 2, 8, 0}. When neighbouring digits are drawn, we reassign them to these digits with {9%, 7%, 5%, 12%} chance respectively. For example if a trailing digit of 4 or 6 is drawn, it is reassigned to a 5 with 9% chance. The imposed rates are similar to those found empirically at heavily biased stations (Nese, 1994), where the cause of the bias is thought to have been introduced through human preference for rounding to certain digits. The introduction of these biases yields similarly low misclassification rates to the unbiased case, with error rate distributions statistically indistinguishable from those shown in Figure 5.

In a more stringent test we also examine the misclassification rate when the length of the sequences varies. We generate 10-year-long sequences using one precision, to which we append a shorter sequence using a second precision. Misclassification rates are highly state-dependent when the second subsequence is short compared with the first (Figure 6), with nearly immediate resolution of brief switches between some pairs of states, but up to $\mathcal{O}(100)$ days being needed for the 0.1 °F-to-0.1 °C case. This limitation arises during the model selection procedure, which

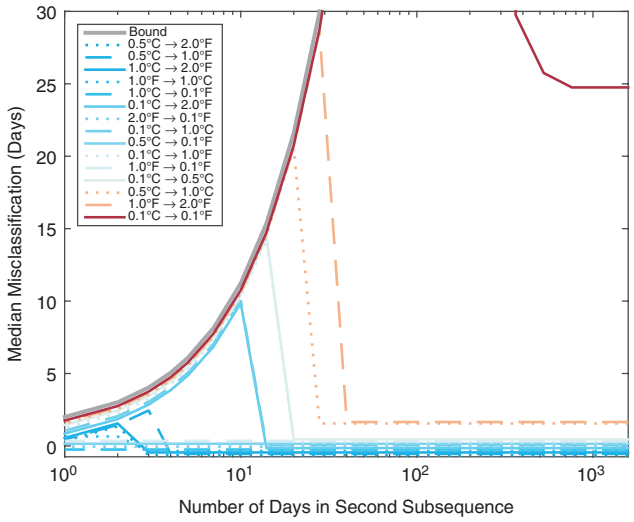

Figure 6. Sensitivity to small number of samples at a second precision. The median number of misclassified days is plotted for precision-decoded synthetic time series having 20 years at one precision, followed by a variable-length sequence at a second precision, as indicated by the x-axis. The 0.1 °F-to-0.1 °C curve extends beyond the visible range of the y-axis, falling after $\mathcal{O}(100)$ days. Lines are offset slightly in the vertical for visual clarity. The worst case bound (grey) corresponds to all samples in the second subsequence being misclassified.

requires sufficient evidence of an additional precision being present to permit the additional model complexity. Once samples of the second precision are sufficiently abundant that the model complexity penalty in the BIC (the second term in Eq. (9)) is outweighed by the benefit of including an additional precision state, the misclassification rate typically drops rapidly. This phenomenon is known as the lock-in time. As with the constant sequence length case (Figure 5), the lock-in time misclassification errors roughly scale with the JSD values (Figure 6). We also repeated the same test but with the first subsequence having the same variable length as the second. When the two subsequences are of equal length, we find lower misclassification rates relative to the unequal length case as a result of the BIC restriction on complexity playing a smaller role.

In general, the misclassification rate is proportional to the number of state switches due to the unavoidable chance of ambiguity for observations immediately before and after transitions. We note that many time series in the GHCND database are $\mathcal{O}(100$ year) in length, leading to expected misclassification rates of $\mathcal{O}(10^{-5})$ for stations that experience one precision transition. Overall, the tests indicate that very small misclassification rates can be expected even when the sampling is moderately biased. Goodness-of-fit tests may reveal stations for which the assumption of unbiased recording does not hold.

In a situation where one has reason to expect regular precision cycles –such as when technicians or instrumentation change seasonally–posterior decoding (Durbin *et al.*, 1998) or related methods may be used to design a test for such simple types of periodicity. Posterior decoding also provides a useful measure of the uncertainty of recovered states. At detected transitions, irreducible ambiguity in the consistency subsets can lead to the posterior probability of the transition location being blurred in time. In plots of the Viterbi path (e.g. Figure 2), we give a visual indication of the 95% confidence interval for the transition (Durbin *et al.*, 1998), although this range is typically narrow (Figures 5 and 6).

4. Results and discussion

Finite precision has a variety of implications for the analysis and interpretation of observations. We discuss how the majority of data are not centred on their most likely value, how this has implications for identification of record-breaking temperatures, and why inference of the original precision is necessary to reliably perform quantile regression. The analysis presented here is by no means an exhaustive study of precision variability within the GHCND database, and precision-decoding will be most useful when used in tandem with other quality-control tools.

4.1. *GHCND data*

We use precision-decoding to estimate the precision of all temperature observations in the public version of the GHCND database, comprising 28 824 weather stations with a total of 644 million samples, and find that 63% of the data are misaligned due to double-rounding. We note that while double-rounding errors are small compared with estimates of the accuracy of surface station temperatures ($\sim 0.5-1.0$ °C; Menne *et al.*, 2012), the inferred precision is generally similar in magnitude. The mean absolute error across the temperature observations in GHCND is 0.018 °C. That so many observations are affected arises from the high density of stations in the USA using constant 1 °F precision, including 1218 US Historical Climate Network stations that are maintained specifically for long-term climate studies (Menne *et al.*, 2009). In total, 71% of all observations are detected as being 1 °F-precision. One-ninth of the 71% are evenly divisible by the five-ninths conversion factor to Celsius, making $(8/9) \times 71\% \approx 63\%$ of observations misaligned.

The prevalence of 1 °F-precision data (Figure 7) accounts for the dearth of fives in the trailing digit of the rounded Celsius data (Figure 1(a)). Additionally, the binning of the original 1 °F-precision data is at intervals that correspond to 0.56 °C–much larger than the implied 0.1 °C intervals at which the data are archived. The original precision of some archived data is 2 °F (0.3% of samples), corresponding to 1.1 °C intervals (Figure 7). Seven hundred and forty stations apparently use 2 °F-precision for more than 1 month of samples. Celsius measurements with either 0.5 °C (4% of samples) or 1 °C (1% of samples) precision obviously also feature greater uncertainty than suggested by the 0.1 °C archival precision (Figure 8). Twenty-four percent of samples are detected as being at least 0.1 °C-precision. Though difficult to distinguish from 0.1 °C, we infer that 7% of these high-precision data are likely 0.1 °F-precision.

Precision changes are also common within individual station time series, with 15% of stations having fewer than 95% of observations in their most common precision category. The appearance of 0.1 °F-precision observations (Figure 7) stems mostly from Australian stations, consistent with the varying, site-specific practices known to have been used there historically (Torok and Nicholls, 1996). A high density of 0.5 °C observations between 1977 and 2012 is primarily from approximately 2300 stations in Canada and Mexico. One degree Celsius-precision observations also appear between 1972 and 1981 in approximately 1000 stations from the former Soviet Union.

4.2. *Effects of unknown precision on quantile regression*

Many statistical methods assume that data are continuously distributed, and corrections that permit the use of discrete measurements typically require specification of the censoring or rounding process. For example, estimation of conditional quantiles is sensitive to finite-precision sampling because non-smooth distributions of the data lead to objective functions that are not differentiable (Machado and Silva, 2005). If non-smoothness is not mitigated, this problem precludes the use of a broad class of statistical methods including quantile regression (Koenker and Bassett, 1978). Remedies include the explicit use of censoring in a regression model (Reich and Smith, 2013) and the application of jitter (i.e. random noise) to approximately restore the original distribution (Machado and Silva, 2005), but both solutions require that the censoring process is known. Whereas adding uniform white noise with a range of ±0.05 °C may otherwise appear sensible given how the data are archived, such an approach would generally not adequately restore GHCND data toward its original distribution and would still give biased results. As shown in the following example, our results permit

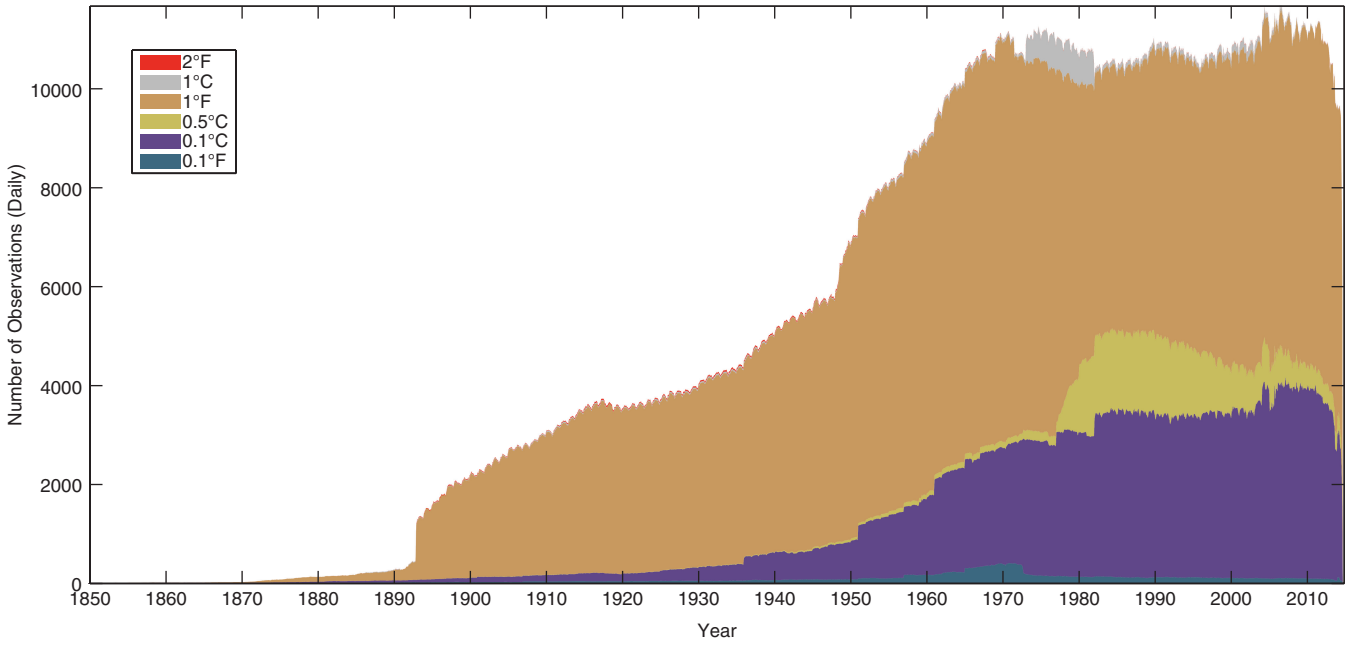

Figure 7. Variations in the distribution of precision and units of daily maximum temperature observations from the GHCND database between 1850 and 2013 detected using precision-decoding. Curves are smoothed in time for the purposes of presentation using a 30-day boxcar window.

accurate restoration of the original underlying data distribution and unbiased quantile regression results.

We use synthetic observations to show that, for rounding procedures common in the GHCND data, these effects are severe even for relatively simple sample distributions. We generate 2000 synthetic time series by drawing from a normal distribution with a standard deviation of 10 °C and adding a linear trend in the mean of 0.02 °C year^{-1}. Each time series contains 90 samples per year over a duration of 30 years, representing daily samples for one season over a climatological period (used, for example, to estimate the base temperature distribution and its trends). Quantile regression is performed to estimate linear trends in the 95th percentile of the time series (Figure 9(a)). These quantile trends are then compared with cases involving realistic sampling practices. In one case, the data are sampled at 1 °C-precision by rounding, and in a second case at 1 °F-precision, followed by re-rounding at 0.1 °C-precision as in the GHCND database.

The same quantile regression procedure is repeated for each time series (Figure 9(b,c)), resulting in significant distortions of the distribution, wherein the inferred trends are strongly biased towards zero. We then show that the application of jitter can restore the distribution, but only if the correct precision is known *a priori*. First, we assume that only the archival precision (0.1 °C for GHCND)—rather than the actual sampling precision—is known. Uniform random values generated on the interval $[0.05, 0.05]$ °C are added to all samples, and the 95th percentile trends are re-estimated (Figure 9(d,e)), yielding estimates that are still severely distorted and zero-inflated with respect to the true distribution. Second, we use the actual sampling precision, as would be available from precision-decoding. For the 1 °C-precision samples, uniform random values on the interval $[-0.5, 0.5]$ °C are added to the time series. Similarly, random values on $[-0.5, 0.5]$ °F are added to each of the 1 °F-precision samples. Again, the 95th percentile slopes are re-estimated (Figure 9(f,g)), yielding a distribution of trends that is approximately restored. A paired Kolmogorov–Smirnov test rejects the null hypothesis of the trends in Figure 9(b)–(e) being drawn from the same distribution as those in (a), while failing to reject the null in the case of (f) and (g) at $p < 0.05$.

In addition to exhibiting zero-inflation and a distorted shape, the means of the uncorrected distributions are biased. The mean of the slope estimates (in °C decade^{-1}) is 0.198 for the original data, while for the 1 °C-rounded data it is 0.180 when no jitter is applied, 0.183 for the incorrectly jittered data, and 0.198 for the correctly jittered data. The mean bias depends on the distribution of the original slopes, the length of the time series, the variability of the data, and the level of rounding. Sufficiently coarse rounding can lead to all slopes being zero to within machine precision.

4.3. *Identification of record-breaking values*

If not corrected, double-rounding can induce errors in the attribution of record events. For example, although no recording practice changes are indicated in the station metadata or in the original source data available from the International Surface Temperature Initiative (Thorne *et al.*, 2011), we find that a Kenyan station (KE000063820, Figure 2(b)) used 1 °F-precision from 1955 to 1961, 1 °C from 1973 to 1981, and finally 0.1 °C from 1982 to present. Correcting the misaligned 1 °F-precision values for daily maxima at this station indicates that 21 of the current maximum records (i.e. each corresponding to a particular day of the year) are misaligned in the GHCND database. For instance, the raw GHCND record at station KE000063820 for 30 May is 30.6 °C in 1961, a record that was tied in 2004. However, we infer that it was originally recorded as 87 °F and was biased upwards by double-rounding after conversion to Celsius. The record for 30 May is therefore actually more likely to have occurred in 2004, when it was recorded with no bias as 30.6 °C.

Chances of misidentifying record events increase with the probability that any two randomly chosen samples from a given station are recorded using different precisions. The probability of finding different precisions can be quantified using a diversity index. We use the Gini–Simpson index (Jost, 2006),

$$G = 1 - \sum_i p_i^2, \tag{13}$$

where $P = \{p_1, p_2, \ldots, p_R\}$ is the categorical distribution of precision states, giving the likelihood that any two randomly chosen samples from a given station will have been recorded using different precisions. Maps of the diversity index (Figure 8) show high probabilities in regions that have switched recording conventions, such as in Canada, where recording conventions for many stations switched from Fahrenheit to Celsius in the 1970s, and Eastern Europe, where precision levels have been variable.

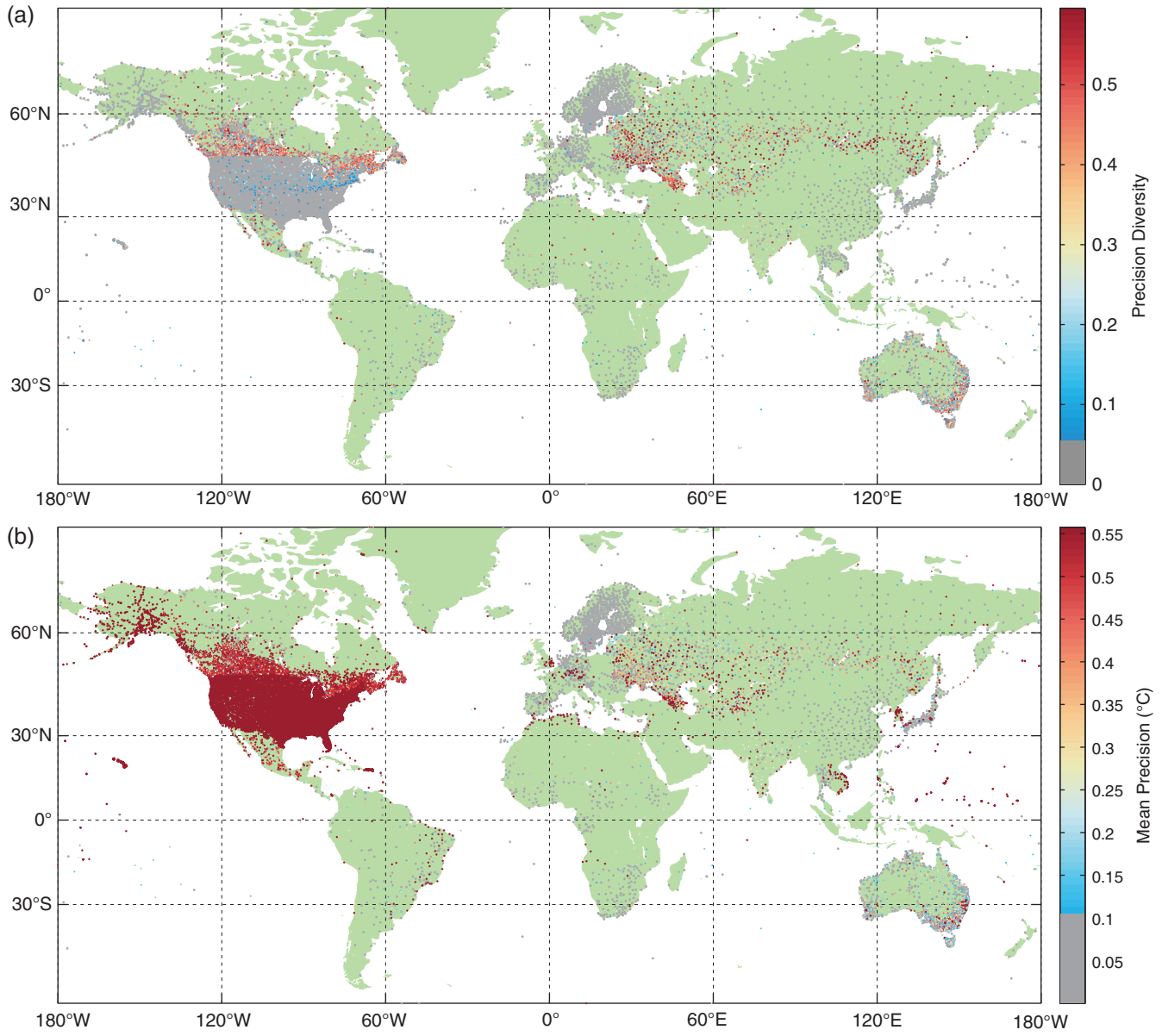

Figure 8. Geographical patterns of precision and its variability. (a) Distinct geographical patterns emerge in a map of precision diversity (Eq. (13)), the probability that any two randomly chosen samples are of different precisions. (b) The mean precision (the average width of the preimage of each measurement). Stations with diversity indices less than 0.05 and mean precisions less than 0.105,°C are displayed in (a) and (b), respectively, as grey points in the background.

An arguably safer approach to identifying record temperatures is to evaluate them probabilistically within the range of each observation's precision, rather than by the daily best-estimate provided by the raw observation. Such an analysis could be performed using Monte-Carlo techniques given that the analytical treatment using beta distributions becomes unwieldy after a small number of candidate record events.

4.4. Improved quality control

In addition to correcting for distributional biases, precision-decoding is a useful tool for quality control. Surface station time series which have been digitized, perhaps after previous aggregation steps, may in some cases represent distinct stations or instruments that have been merged because of the appearance of continuity. For example, a Russian station (RSM00027037, Figure 2(a)) contains several years of 1 °F observations at the start of the record, followed by 0.1 °C for the remainder. The source flags within the data indicate that the early observations were transcribed from global radio weather bulletins by operators in the USA, while the later, more precise, data are sourced directly from a Russian database (Razuvaev *et al.*, 2008), consistent with the detected switch. In many cases, however, source flags do not provide such a clear transcript, and any detected precision switches indicate that undocumented changes may have occurred. This information can be used to improve the power of change-point detection tests currently used in homogenization and quality-control procedures (Karl and Williams, 1987; Lund and Reeves, 2002; Menne and Williams, 2009).

In another case, it appears that 528 US stations located at municipal airports reported unusual below-freezing temperatures. The values are clearly spaced at intervals of 1 °F, but with an offset of 0.1 °C from any plausible sequence of conversion and rounding. The offset leads to these samples being detected as transitions from 1 °F to either 0.1 °C or 0.1 °F precision. For example, since 2006 station USW00003866 has reported 363 temperature minima and eight maxima ranging from −13.8 to −0.6 °C that, after conversion from Fahrenheit, seem to have been rounded upwards instead of to the nearest 0.1 °C. The fact that these errors become apparent only in the 2000s and that they occur only for sub-freezing temperatures is suggestive of a software error or some other change in data processing or instrumentation.

5. Conclusion

Statistical methods, especially those concerned with assessing distributional changes or temperature extremes on daily

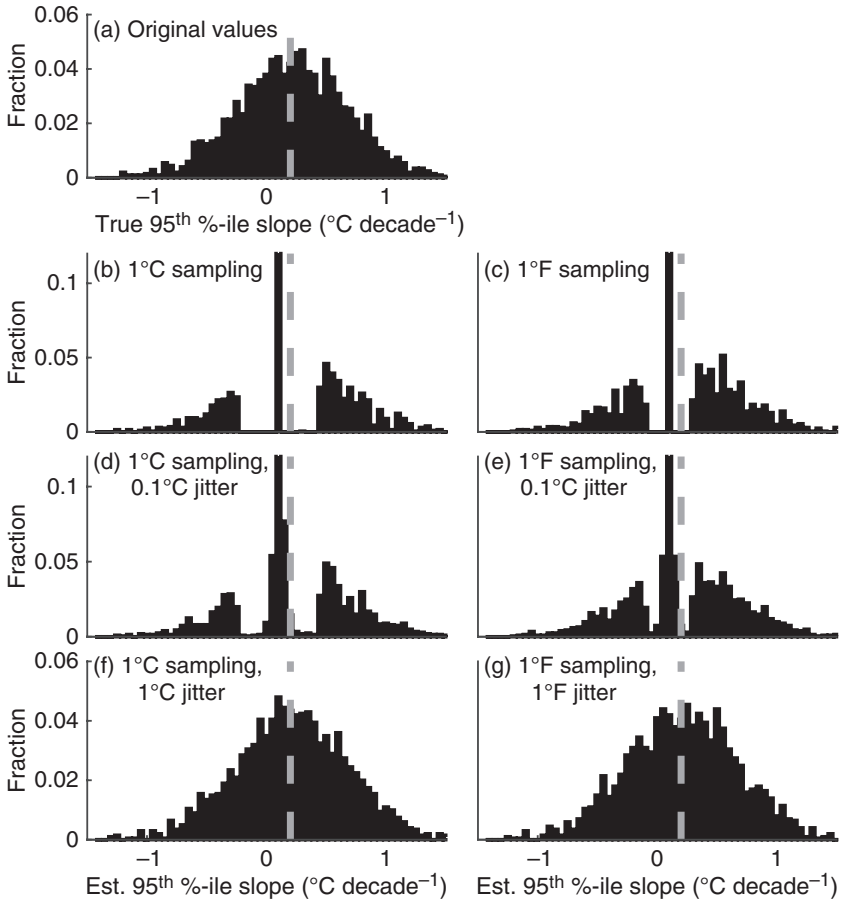

Figure 9. The effect of finite-precision sampling on quantile regression. Quantile regression is performed to estimate trends in the 95th percentile of synthetic temperature time series with a prescribed distribution of linear trends in the mean (a). The analysis is repeated for the same time series sampled at (b) 1 °C and (c) 1 °F precision, with significant zero-inflation evident due to non-smoothness of the data. When jitter of the same 0.1 °C magnitude as the GHCND archival precision is added, the distributions for (d) 1 °C and (e) 1 °F precision are still highly distorted. When the correct (f) 1 °C and (g) 1 °F magnitude jitter is added to the rounded data, the original distribution is restored. In all panels, the prescribed mean trend of 0.02 °C year^{-1} is indicated by the vertical dashed line.

time-scales, are sensitive to rounding, double-rounding, and precision or unit changes. The precision-decoding algorithm presented here provides an efficient and robust method of inferring precision of time series and permits for correcting some of the biases or errors which would otherwise accrue. Application of precision-decoding to the GHCND database shows that 63% of all temperature observations are misaligned due to unit conversion and double-rounding, and that many time series contain substantial changes in precision over time.

The effects of double-rounding do not lead to systematic biases in quantities that average over many temperature observations because individual biases are asymptotically symmetric about the true value when variability substantially exceeds the precision. For quantities such as global mean temperature or monthly averages at a single station, variable precision does not in itself lead to significantly greater uncertainty. For other statistics, however, substantial sensitivity is demonstrated. Methods that rely on either distributional smoothness or individual extreme values are shown to be particularly vulnerable, but generally amenable to bias-correction through precision-decoding.

It may prove useful to combine or augment precision-decoding with additional methods which can account for the available metadata or other prior information. For example, a Bayesian approach could be implemented to make use of geographical or station network information in weighting the different sets of precision candidates prior to model selection. Existing Markov methods such as pair-HMMs (Durbin *et al.*, 1998) used for sequence alignment would also be a powerful means of pooling precision information across a station network, though algorithmic performance would currently limit their applicability to small subsets of stations.

A software implementation of the precision-decoding algorithm and the inferred precision of all publicly available GHCND temperature station data is available in an online database (Rhines, 2015). We expect that these results will be useful in further error detection, change-point analyses, and examinations of station temperature distributions, amongst other applications. Precision-decoding can also be applied to other temperature data such as sea-surface temperature observations (Worley *et al.*, 2005) and vertical temperature profiles from radiosonde archives (Durre *et al.*, 2006). Historical observations of variables with physical units other than temperature (including surface pressure, precipitation, and wind speed and direction) provide important constraints for reanalyses (e.g. Compo *et al.*, 2006) and other datasets, and precision-decoding may be useful for additional quality control there as well. We note that measurements involving length units such as surface pressure, precipitation, and sea level (Woodworth and Player, 2003; Brohan *et al.*, 2009) are also susceptible to the double-rounding issues discussed here because of historical changes between the Imperial and metric systems.

Acknowledgements

This work was funded by NSF P2C2 grant 1304309. Data analysis was performed on the Odyssey cluster supported by the FAS Science Division Research Computing Group at Harvard University.

References

Al-Marzouki S. 2005. Are these data real? Statistical methods for the detection of data fabrication in clinical trials. *Br. Med. J.* **331**: 267–270, doi: 10.1136/bmj.331.7511.267.

Austin JF, McConnell A. 1980. Two hundred years of the Six's self-registering thermometer. *Notes Rec. R. Soc.* **35**: 49–65, doi: 10.1098/rsnr.1980.0004.

Bogdan M, Ghosh JK, Zak-Szatkowska M. 2008. Selecting explanatory variables with the modified version of the Bayesian information criterion. *Qual. Reliab. Eng. Int.* **24**: 627–641, doi: 10.1002/qre.936.

Brohan P, Kennedy JJ, Harris I, Tett SFB, Jones PD. 2006. Uncertainty estimates in regional and global observed temperature changes: A new data set from 1850. *J. Geophys. Res.* **111**: D12106, doi: 10.1029/2005jd006548.

Brohan P, Allan R, Freeman JE, Waple AM, Wheeler D, Wilkinson C, Woodruff S. 2009. Marine observations of old weather. *Bull. Am. Meteorol. Soc.* **90**: 219–230, doi: 10.1175/2008bams2522.1.

Compo GP, Whitaker JS, Sardeshmukh PD. 2006. Feasibility of a 100-year reanalysis using only surface pressure data. *Bull. Am. Meteorol. Soc.* **87**: 175–190.

Dempster AP, Laird NM, Rubin DB. 1977. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Statist. Soc. B* **39**: 1–38.

Durbin R, Eddy SR, Krogh A, Mitchison G. 1998. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press: Cambridge, UK.

Durre I, Vose RS, Wuertz DB. 2006. Overview of the integrated global radiosonde archive. *J. Clim.* **19**: 53–68, doi: 10.1175/JCLI3594.1.

Durre I, Menne MJ, Gleason BE, Houston TG, Vose RS. 2010. Comprehensive automated quality assurance of daily surface observations. *J. Appl. Meteorol.* **49**: 1615–1633, doi: 10.1175/2010JAMC2375.1.

Durtschi C, Hillison W, Pacini C. 2004. The effective use of Benford's law to assist in detecting fraud in accounting data. *J. Forensic Acc.* **5**: 17–34.

Feng S, Hu Q, Qian W. 2004. Quality control of daily meteorological data in China, 1951–2000: A new dataset. *Int. J. Climatol.* **24**: 853–870, doi: 10.1002/joc.1047.

Figueroa SA. 1995. When is double rounding innocuous?. *ACM SIGNUM Newsletter* **30**: 21–26.

Forney GD, Jr. 1973. The Viterbi algorithm. *Proc. IEEE* **61**: 268–278.

Jost L. 2006. Entropy and diversity. *Oikos* **113**: 363–375, doi: 10.1111/j.2006.0030-1299.14714.x.

Karl TR, Williams CN. 1987. An approach to adjusting climatological time series for discontinuous inhomogeneities. *J. Clim. Appl. Meteorol.* **26**: 1744–1763, doi: 10.1175/1520-0450(1987)026<1744:aatact>2.0.co;2.

Koenker R, Bassett G, Jr. 1978. Regression quantiles. *Econometrica* **46**: 33–50.

Lin J. 1991. Divergence measures based on the Shannon entropy. *IEEE Trans. Inf. Theory* **37**: 145–151, doi: 10.1109/18.61115.

Lund R, Reeves J. 2002. Detection of undocumented changepoints: A revision of the two-phase regression model. *J. Clim.* **15**: 2547–2554, doi: 10.1175/1520-0442(2002)015<2547:doucar>2.0.CO;2.

Machado JAF, Silva JS. 2005. Quantiles for counts. *J. Am. Stat. Assoc.* **100**: 1226–1237, doi: 10.1198/016214505000000330.

Menne MJ, Williams CN, Jr. 2009. Homogenization of temperature series via pairwise comparisons. *J. Clim.* **22**: 1700–1717, doi: 10.1175/2008JCLI2263.1.

Menne MJ, Williams CN, Vose RS. 2009. The US historical climatology network monthly temperature data, version 2. *Bull. Am. Meteorol. Soc.* **90**: 993–1007, doi: 10.1175/2008bams2613.1.

Menne MJ, Durre I, Vose RS, Gleason BE, Houston TG. 2012. An overview of the global historical climatology network-daily database. *J. Atmos. Ocean Tech.* **29**: 897–910, doi: 10.1175/JTECH-D-11-00103.1.

Muller RA, Wurtele J, Rohde R, Jacobsen R, Perlmutter S. 2013. Earth atmospheric land surface temperature and station quality in the contiguous United States. *Geoinfor. Geostat.* **1**: 1–6, doi: 10.4172/2327-4581.1000107.

Nese JM. 1994. Systematic biases in manual observations of daily maximum and minimum temperature. *J. Clim.* **7**: 834–842, doi: 10.1175/1520-0442(1994)007<0834:sbimoo>2.0.CO;2.

Preece DA. 1981. Distributions of final digits in data. *J. R. Stat. Soc.* **30**: 31–60.

Rabiner LR. 1989. A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE* **77**: 257–286, doi: 10.1109/5.18626.

Razuvaev VN, Apasova EG, Martuganov RA. 2008. 'Daily temperature and precipitation data for 223 former-USSR stations', Technical report ORNL/CDIAC-56, NDP-040. Carbon Dioxide Information Analysis Center, Oak Ridge National Laboratory, US Department of Energy: Oak Ridge, TN, doi: 10.3334/CDIAC/cli.ndp040.

Reek T, Doty SR, Owen TW. 1992. A deterministic approach to the validation of historical daily temperature and precipitation data from the cooperative network. *Bull. Am. Meteorol. Soc.* **73**: 753–762, doi: 10.1175/1520-0477(1992)073<0753:adattv>2.0.CO;2.

Reich BJ, Smith LB. 2013. Bayesian quantile regression for censored data. *Biometrics* **69**: 651–660, doi: 10.1111/biom.12053.

Rhines A. 2015. 'A MATLAB implementation of the precision-decoding algorithm and estimated precision for the GHCND'. http://www.stochtastic.com/ghcnd.html (accessed 20 July 2015).

Schwarz G. 1978. Estimating the dimension of a model. *Ann. Statist.* **6**: 461–464, doi: 10.1214/aos/1176344136.

Thorne PW, Willett KM, Allan RJ, Bojinski S, Christy JR, Fox N, Gilbert S, Jolliffe I, Kennedy JJ, Kent E. 2011. Guiding the creation of a comprehensive surface temperature resource for twenty-first-century climate science. *Bull. Am. Meteorol. Soc.* **92**: ES40–ES47, doi: 10.1175/2011bams3124.1.

Torok S, Nicholls N. 1996. A historical annual temperature dataset for Australia. *Aust. Meteorol. Mag.* **45**: 251–260.

Wang K. 2014. Sampling biases in datasets of historical mean air temperature over land. *Nature Sci. Rep.* **4**: 4637, doi: 10.1038/srep04637.

Woodworth PL, Player R. 2003. The permanent service for mean sea level: An update to the 21st century. *J. Coastal Res.* **19**: 287–295.

Worley SJ, Woodruff SD, Reynolds RW, Lubker SJ, Lott N. 2005. ICOADS release 2.1 data and products. *Int. J. Climatol.* **25**: 823–842, doi: 10.1002/joc.1166.

Zhang X, Zwiers FW, Hegerl G. 2009. The influences of data precision on the calculation of temperature percentile indices. *Int. J. Climatol.* **29**: 321–327, doi: 10.1002/joc.1738.

