https://doi.org/10.5194/amt-2020-454
Preprint. Discussion started: 4 January 2021

Atmospheric Measurement Techniques
Discussions
Open Access
EGU

Global Ensemble of Temperatures over 1850-2018: Quantification of Uncertainties in Observations, Coverage, and Spatial modelling (GETQUOCS)

Maryam Ilyas[a,b], Douglas Nychka[c], Chris Brierley[d], and Serge Guillas[a]

[a]Department of Statistical Science, University College London, London, UK
[b]College of Statistical and Actuarial Sciences, University of the Punjab, Lahore, Pakistan
[c]Applied Mathematics and Statistics, Colorado School of Mines, Golden, CO, USA
[d]Department of Geography, University College London, London, UK

Correspondence: maryam.stat@pu.edu.pk, maryam.ilyas.14@ucl.ac.uk

Abstract. Instrumental temperature records are derived from the network of in situ measurements of land and sea surface temperatures. This observational evidence is seen as fundamental to climate science. Therefore, the accuracy of these measurements is of prime importance for the analysis of temperature variability. There are spatial gaps in the distribution of instrumental temperature measurements across the globe. This lack of spatial coverage introduces coverage error. An approximate Bayesian computation based multi-resolution lattice kriging is developed and used to quantify the coverage errors through the variance of the spatial process at multiple spatial scales. It critically accounts for the variation in the parameters of this advanced spatial statistics model itself, thereby providing for the first time a full description of both the spatial coverage uncertainties along with the uncertainties in the modelling of these spatial gaps. These coverage errors are combined with the existing estimates of uncertainties due to observational issues at each station location. It results in an ensemble of 100,000 monthly temperatures fields over the entire globe that samples the combination of coverage, parametric and observational uncertainties from 1850 till 2018 over a 5° × 5° grid.

1 Introduction

Instrumental surface temperature data sets are frequently used to determine natural variability of changing surface temperatures on Earth (e.g. Hansen et al., 2010; Morice et al., 2012; Good, 2016; Menne et al., 2018). Climate models also use instrumental observations for accurate assessment of various climate phenomenon (e.g. Milton and Earnshaw, 2007; Edwards et al., 2011; Glanemann et al., 2020). Temperature data bases are generally created by blending the land and sea surface temperature records. The land component of the data sets is mostly collected from the global historical network of meteorological stations (e.g. Jones et al., 2012). These are obtained from the World Meteorological Organization (WMO) and Global Climate Observation System (GCOS). On the other hand, sea surface temperatures are largely compiled by the International Comprehensive Ocean-Atmosphere Data Set (ICOADS) (Woodruff et al., 2011). These are collected from ships and drifting buoys (e.g.

Kennedy et al., 2011b).

These raw temperature estimates are post-processed by removing biases from them (Dunn et al., 2014). In a first step of quality control, noise originating from instrumental or observer error is removed (Dunn et al., 2016). After this, systematic biases that arise from station movements or incorrect station merges, changes in instruments and observing practices and land use changes around stations (more commonly known as urbanization impacts) are removed (Dunn et al., 2016). Such a homogenization process (Domonkos and Coll, 2017) aims to remove or at least reduce the non-climatic signals that will likely affect the genuine data characteristics (e.g. Hausfather et al., 2016; Cao et al., 2017).

Blended land and sea surface temperature data are generated by a variety of organizations. These include: Global Surface Temperatures (MLOST) by the National Oceanic and Atmospheric Administration (NOAA) (Smith et al., 2008; Vose et al., 2012), Goddard Institute for Space Studies (GISS) surface temperature anomalies by the National Aeronautics and Space Administration (NASA) (Hansen et al., 2010; Lenssen et al., 2019), temperature anomalies provided by Japanese Meteorological Agency (JMA) (Ishii et al., 2005), HadCRUT temperature anomalies by the Met Office Hadley Centre and the University of East Anglia Climatic Research Unit (Morice et al., 2012), and Berkeley Earth Surface Temperature (BEST) by Rhode et al. (2013). Each group compiles these monthly temperature products using somewhat different input data, and extensively different quality control and homogenization procedures (e.g. Rohde, 2013; Jones, 2016).

GISS makes substantial use of satellite data (Hansen et al., 2010); MLOST only uses satellite data in a limited capacity (Smith et al., 2008); and HadCRUT and BEST use no satellite data at all (Morice et al., 2012; Rhode et al., 2013). These data sets also are different in terms of their starting years: 1850-present for HadCRUT and BEST; 1880-present for GISS and NOAA; and 1891-present for JMA. The spatial resolution is different as well. Each group also employs different methods of averaging to derive gridded temperature products from in situ measurements (Jones, 2016; McKinnon et al., 2017).

In addition to these methodological differences, spatial coverage is also being treated differently by these groups (Huang et al., 2020). The HadCRUT4 and JMA datasets do not interpolate over grid boxes having missing observations after the gridding stage. However, the sea component of JMA grid estimates are based on optimally interpolated (i.e. kriging) sea surface temperature anomalies (Ishii et al., 2005; Kennedy, 2014). On the other hand, no spatial interpolation is performed on HadSST3 (Rayner et al., 2006) and CRUTEM4 (Jones et al., 2012) that are the land and sea components of HadCRUT4 data set. The MLOST performs linear spatial interpolation using nearby stations in areas lacking stations (Smith et al., 2008). For broader spatial coverage, the GISS uses a linear inverse distance weighting with data from all the stations up to 1200 km of the prediction location (Hansen et al., 2010). The weight of each sample point decreases linearly from unity to zero. This interpolation scheme computes estimates by weighting the sample points closer to the prediction location greater than those farther away without considering the degree of autocorrelation for those distances. On the other hand, the JMA (**?**) records use covariance structure of spatial data and are based on traditional kriging. Formal Gaussian process regression is used by the BEST to produce spatially complete temperature estimates (Rhode et al., 2013). Cowtan and Robert (2014) also handle the the issue of missing observations and provide a data product that is based on HadCRUT4 temperature estimates (Morice et al., 2012). This data set (Cowtan and Robert, 2014) consists of spatially dense fields. The unobserved grid cells of HadCRUT4 spatial fields are estimated using a spatial interpolation approach i.e. ordinary kriging. It is worth noting that all these interpolation approaches ignore the variations at multiple scales that exist in the climate system. Additionally, regional coverage uncertainty estimates are not available for these temperature data sets.

Recently, a new monthly temperature data set was created (Ilyas et al., 2017). It employs the multi-resolution lattice kriging approach (Nychka et al., 2015) that captures variation at multiple scales of the spatial process. This multi-resolution model quantifies gridded uncertainties in global temperatures due to the gaps in spatial coverage. It results in a 10,000 member ensemble of monthly temperatures over the entire globe. These are spatially dense equally plausible fields that sample the combination of observational and coverage uncertainties. The data are for open access and freely available at: https://oasishub.co/dataset/global-monthly-temperature-ensemble-1850-to-2016.

This paper provides a substantial update on Ilyas et al. (2017) data set. Here, a new version of this data set is produced that incorporates the uncertainties in the statistical modelling itself (i.e. parametric uncertainties) in addition to the observational and coverage errors. Hence, this data set can be considered to be the first to quantify all types of uncertainties in past global temperatures, including the statistical model's inference process. To account for the model parametric uncertainties, an approximate Bayesian inference methodology is proposed that extend the multi-resolution lattice kriging (Nychka et al., 2015). It is based on the variogram, a measure of spatial variability between spatial observations as a function of spatial distance.

2 Methods

2.1 Multi-resolution lattice kriging using ABC

The Multi-resolution lattice kriging (MRLK) model was introduced by Nychka et al. (2015). It models spatial observations as a sum of a Gaussian process, a linear trend and a measurement error term. The MRLK can flexibly adjust to complicated shapes of the spatial domain and has the property of approximating standard covariance functions. This methodology extends spatial methods to very large data sets accounting for all scales, for the goals of spatial inference and prediction. Indeed, it is computationally efficient for large data sets by exploiting sparsity in covariance matrices. The underlying spatial process is a sum of independent processes, each of which is a linear combination of the chosen basis functions. The basis functions are fixed and coefficients of the basis functions are random.

Consider observations $y(\mathbf{x})$ at n spatial locations $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n$ in the spatial domain D. The aim is to predict the underlying process at an arbitrary location $\mathbf{x} \in D$ and to estimate the uncertainty in the prediction. For $\mathbf{x} \in D$,

$$y(\mathbf{x}) = d + g(\mathbf{x}) + \epsilon(\mathbf{x}) \tag{1}$$

where d is the mean and ϵ the error term. The unknown spatial process $g(\mathbf{x})$ is assumed to be the sum of L independent processes having different scales of spatial dependence. Each process is a linear combination of m basis functions where $m(l)$ is the number of basis functions at level l,

$$g(\mathbf{x}) = \sum_{l=1}^{L} g_l(\mathbf{x}) = \sum_{l=1}^{L} \sum_{j=1}^{m(l)} c_j^l \phi_{j,l}(\mathbf{x}) \tag{2}$$

The basis functions ($\phi_{j,l}$) are fixed. These are constructed at each level using the unimodal and symmetric radial basis functions. Radial basis functions are functions that depend only on the distance from the center.

The inference methodology (Nychka et al., 2015) is the direct consequence of maximizing the likelihood function. This inference framework does not account for the uncertainty in the model parameters within the MRLK (Nychka et al., 2015). Here, we estimate the MRLK parameters and quantify uncertainty in these parameters. For this purpose, a Bayesian framework is created in which the posterior densities of the multi-resolution lattice kriging parameters are estimated using the Approximate Bayesian Computation (ABC). Our new technique allows for the spatial predictions to be accompanied by a quantification of uncertainties in these predictions that reflect not only the coverage gaps but also the uncertainties in the MRLK parameters.

2.1.1 ABC posterior density estimation

Consider a n-dimensional spatial random variable $\mathbf{y}(\mathbf{x})$. The multi-resolution lattice kriging model depends on the unknown p-dimensional parameter $\boldsymbol{\theta}$. The probability distribution of the data given a specific parameter value $\boldsymbol{\theta}$ is denoted by $f(\mathbf{y}|\boldsymbol{\theta})$. If the prior distribution of $\boldsymbol{\theta}$ is denoted as $\pi(\boldsymbol{\theta})$, then the posterior density is given by

$$f(\boldsymbol{\theta}|\mathbf{y}) \propto f(\mathbf{y}|\boldsymbol{\theta})\, \pi(\boldsymbol{\theta}) \tag{3}$$

Here, $\boldsymbol{\theta} = [\lambda, aw]^T$ where λ and aw are respectively the smoothing parameter and autoregressive weights. These are the two main parameters of the MRLK. The autoregressive weight aw is the key covariance parameter. It is essential for specifying and fitting the spatial model. The smoothness parameter λ influences throughout the calculation: an inappropriate estimate can lead to over or under fitting a spatial model and can result in imprecise interpolated values and prediction uncertainties (Nychka et al., 2015). The posterior distribution of these parameters given data, $f(\boldsymbol{\theta}|\mathbf{y})$, is approximated using ABC. The ABC acceptance-rejection technique based on variogram as a summary statistic is developed in the next section that is used to approximate the posterior densities.

2.1.2 Variogram-based ABC algorithm

Approximate Bayesian computation (e.g. Busetto and Buhmann, 2009; Beaumont, 2010; Dutta et al., 2017; Beaumont, 2019) is a family of algorithms that deals with the situations where the likelihood of a statistical model is intractable, whereas it is possible to simulate data from the model for a given parameter value. ABC bypasses the evaluation of the likelihood function by comparing observed and simulated data. Additionally, it offers algorithms that are very easy to parallelize. There are several forms of ABC algorithms. The standard rejection algorithm is the classical ABC sampler (e.g. Pritchard et al., 1999; Beaumont et al., 2002). It is widely used e.g. for model calibration by Gosling et al. (2018). The algorithm is based on drawing values of the parameters from the prior distribution. The data sets are simulated for each draw of parameters, each resulting in a chosen summary statistic. A distance metric is computed between the summary statistic of the observed and simulated data. The parameters that produce distances less than a tolerance threshold are retained. These accepted parameters form a sample from the approximate posterior distribution.

The basic idea of ABC is to simulate from the multi-resolution lattice kriging model for a given set of parameters $\boldsymbol{\theta}$. Simulations are run for a large number of parameters to be able to produce meaningful posterior distributions. The parameter values are retained for simulated data $\mathbf{y}^*$ that match the observed data $\mathbf{y}$ up to a tolerance threshold. For the similarity metric, we choose the sum of the squared differences between the semivariance at various lag distances of observed ($\gamma(h)$) and simulated ($\gamma^*(h)$) data. Indeed, these semivariances are traditional descriptors of the correlations across space. The retained $\boldsymbol{\theta*} = [\lambda^*, aw^*]^T$ are such that $\boldsymbol{\theta*} \sim f_t(\boldsymbol{\theta}|\gamma)$.

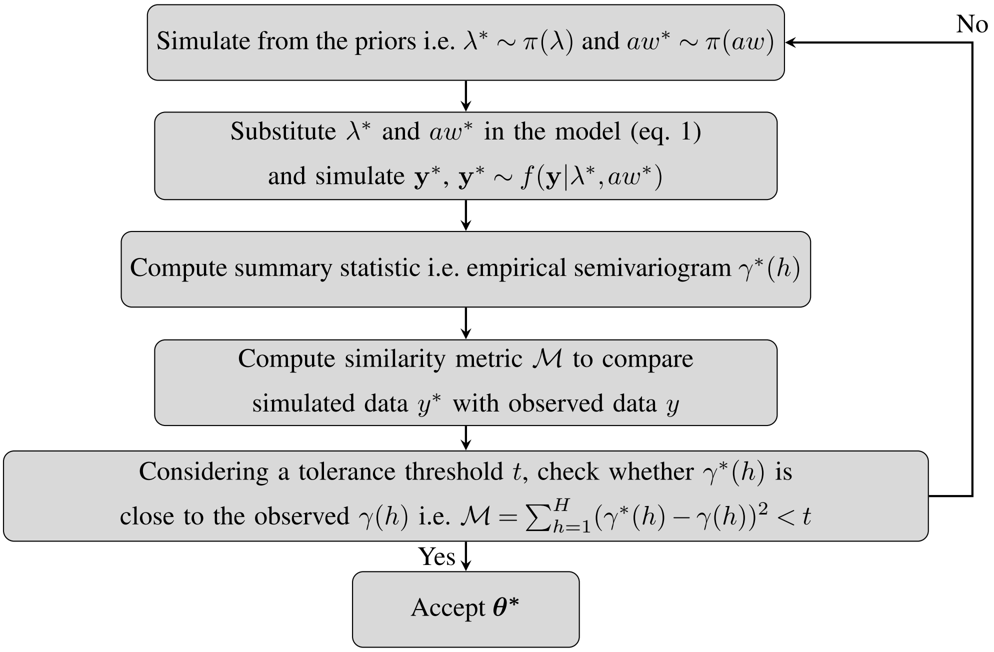
No
Simulate from the priors i.e. $\lambda^* \sim \pi(\lambda)$ and $aw^* \sim \pi(aw)$
Substitute λ^* and aw^* in the model (eq. 1) and simulate $\mathbf{y}^*$, $\mathbf{y}^* \sim f(\mathbf{y}|\lambda^*, aw^*)$
Compute summary statistic i.e. empirical semivariogram $\gamma^*(h)$
Compute similarity metric $\mathcal{M}$ to compare simulated data y^* with observed data y
Considering a tolerance threshold t, check whether $\gamma^*(h)$ is close to the observed $\gamma(h)$ i.e. $\mathcal{M} = \sum_{h=1}^{H}(\gamma^*(h) - \gamma(h))^2 < t$
Yes
Accept $\boldsymbol{\theta*}$

Figure 1. ABC acceptance-rejection algorithm using variogram as summary statistic for multi-resolution lattice kriging

2.2 Spatial estimate based on ABC posteriors

The existing MRLK inference methodology (Nychka et al., 2015) uses maximum likelihood estimators of smoothing parameter λ and autoregressive weights aw. The ABC based inference methodology proposed here accounts for the posterior densities (not joint pointwise estimates) while making spatial predictions. For the full spatial estimate, the ABC posterior distributions

https://doi.org/10.5194/amt-2020-454
Preprint. Discussion started: 4 January 2021

of the MRLK model parameters ($\boldsymbol{\theta} = [\lambda, aw]^T$) are used to determine the conditional distribution of the coefficient $\mathbf{c}$ in (2). For the k$^{\text{th}}$ posterior sample,

$$[\mathbf{c}_k|\mathbf{y}, d, \rho, \theta_k] \sim N_m(\mu_{\mathbf{c}_k}, \boldsymbol{\Sigma}_{\mathbf{c}_k}) \tag{4}$$

However, d and ρ are still estimated using the maximum likelihood approach. Hence, the spatial predictions in ABC based multi-resolution lattice kriging are carried out using (5) and the associated uncertainties are evaluated based on (6) below:

$$\hat{\mathbf{y}} = \frac{\sum_{k=1}^{K}(\mathbf{1}_{(n)}\hat{d} + \boldsymbol{\Phi}\hat{\mu}_{\mathbf{c}_k})}{K} \tag{5}$$

$$\text{Var}(\hat{\mathbf{y}}) = \frac{\sum_{k=1}^{K}(\boldsymbol{\Phi}\hat{\boldsymbol{\Sigma}}_{\mathbf{c}_k}\boldsymbol{\Phi}^T)}{K} \tag{6}$$

3 HadCRUT4 data

The primary data used in this paper are the well-known HadCRUT4 (version 4.5.0.0) temperature anomalies (Morice et al., 2012). It is a combination of global land surface air temperature (CRUTEM4) (Jones et al., 2012) and sea surface monthly temperatures (HadSST3) (Kennedy et al., 2011a, b; Kennedy, 2014). The HadCRUT4 database consists of temperature anomalies with respect to the baseline (1961-1990). Monthly temperatures are provided beginning from 1850 over a 5° × 5° grid. The average temperature anomalies of the stations falling within each grid are provided (Morice et al., 2012). The data set is updated on monthly basis to provide the updated climatic state. The gridded temperature estimates and time series can be downloaded from the Met Office website https://www.metoffice.gov.uk/hadobs/hadcrut4/. It is important to note that a newer and more refined version (HadCRUT5) of this data set is recently created (Morice et al., 2020).

3.1 Ensemble members

Errors in weather observations can either be random or systematic. They can lead to a complex spatial and temporal correlation structures in the gridded data. An ensemble approach is used by Morice et al. (2012) to represent the observational uncertainties in HadCRUT4 data. The ensemble methodology characterizes the uncertainties that are spatially and temporally correlated. The realizations of an ensemble are typically formed by combining the observed data with multiple realizations drawn from the uncertainty model. This uncertainty model describes spatial and temporal interdependencies. This allows one-to-one blending of 100 realizations of HadSST3 and 100 realizations of CRUTEM4, resulting in 100 realizations of HadCRUT ensemble data. Together these HadCRUT ensemble members represent the distribution of observational uncertainties that arise due to the non-climatic factors.

3.2 Observational uncertainties

Systematic observational errors emerge from the non-climatic factors. The HadCRUT ensemble data are created by blending the sea surface temperature anomalies from HadSST3 (Kennedy et al., 2011b, a; Kennedy, 2014) and land temperature anoma-

https://doi.org/10.5194/amt-2020-454
Preprint. Discussion started: 4 January 2021

lies from CRUTEM4 (Jones et al., 2012). This approach follows the use of ensemble methodology to represent a range of observational uncertainties. HadSST3 is an ensemble data that is based on Rayner et al. (2006) uncertainty model. CRUTEM4 is not available as an ensemble data set (Jones et al., 2012). So CRUTEM4 was converted to an ensemble data by Morice et al. (2012) using Brohan et al. (2006) uncertainty model.

The sea surface temperature anomalies are typically being measured using engine room intake measurements, bucket measurements, and drifting buoys. HadSST3 ensemble is used as the sea component of HadCRUT data. This ensemble is generated by drawing the bias-adjustment realizations for three measurement types. Therefore, the ensemble samples the systematic observational errors in sea surface temperature anomalies (Rayner et al., 2006; Kennedy et al., 2011a; Kennedy, 2014). Below are the components of the error model used to characterize the observational uncertainties in the land measurements of HadCRUT. This error model also generates ensemble version of CRUTEM4 (Morice et al., 2012).

Homogenization adjustment error Systematic biases occur due to changing station locations, measurement time, equipment and methods to calculate monthly averages. Homogenization adjustments are applied to the data to remove these non-climatic signals. These adjustments typically do not fully capture the systematic biases. This residual error is referred as homogenization adjustment error. It is modeled using Gaussian distribution (Brohan et al., 2006; Morice et al., 2012). For station records, a value of homogenization adjustment error is drawn from zero mean Gaussian distribution with a standard deviation of $0.4°$C (Morice et al., 2012).

Climatological error For each station, temperature anomalies are computed with respect to the base period $1961-1990$. Typically, data are not available for all the months in the 30-year climatological period. These missing observations introduce climatological error in the estimates of the base-period. The climatological error is modeled using a Gaussian distribution (Morice et al., 2012). For each station, a sample is drawn for each of the 12 calendar months. These realizations are considered same for all the years in the station record (Morice et al., 2012).

Urbanization bias The urban areas absorb and store more heat than the rural areas since the last few decades. This creates a heating effect that is known as the urban heat effect. The urbanization effect induces warming bias in the temperature records. That is, the global temperature estimates include additional warming due to the urban heat effect (Parker, 2010; Feng et al., 2014). This bias is referred as urbanization bias. In HadCRUT, effects of urbanization are modeled on a global scale instead of considering these effects on measurement stations. For this, a truncated Gaussian distribution is used. The large-scale urbanization bias in temperatures is adjusted for all the years beyond 1900 (Morice et al., 2012). Warming rate is sampled from a Gaussian distribution. If a negative value is drawn, it is set as $0.0°$C. Prior to 1900, the urbanization bias is assumed as $0.0°$C as well (Morice et al., 2012).

Exposure bias It has been observed that bias in temperatures can be introduced due to the station siting and exposure. Changes in instrumentation can broadly be grouped into two broad classes. There were few standards for thermometer exposure or instrument shelters before the 19th century. By the early 20th century, these thatched (or covered) enclosures were largely replaced by free-standing louvered shelters or Stevenson type screens (Trewin, 2010). A Stevenson type screen is a shelter or enclosure that protects meteorological instruments from precipitation and direct heat radiation. However, it allows free circulation of the air. Changes in the thermometers, exposure to the atmosphere and shelters from direct or indirect solar radiation introduces exposure bias in temperatures (Parker, 1994; Moberg et al., 2003). This error in temperatures is modeled on a regional scale in HadCRUT using a Gaussian distribution (Morice et al., 2012). Exposure bias uncertainty takes a value of $0.2°$C and $0.1°$C for the stations that respectively lie within and outside the range of $20°$S - $20°$N latitude band (Morice et al., 2012).

4 Hyperparameter temperature ensemble data

The ensemble temperature data set created by Ilyas et al. (2017) presumed perfect knowledge of multi-resolution lattice kriging covariance parameters. The approximate Bayesian computation based multi-resolution lattice kriging developed in Section 2.1 is applied to the sparse HadCRUT4 ensemble data (Section 3, Morice et al. (2012)). As a result of this, a new 100,000 member ensemble data is created. It is an update to the data set discussed in Ilyas et al. (2017). The key difference between the two data sets is the inference methodology. The updated data set is produced by using the ABC based posterior densities of the multi-resolution lattice kriging covariance parameters whereas the first data set used pointwise estimates obtained via a likelihood approach. The use of posterior distribution of the model parameters creates a data set that accounts for the multi-resolution lattice kriging parametric uncertainties.

4.1 ABC posteriors and model parameters

The HadCRUT ensemble data set samples observational uncertainties in the instrumental temperature records (Morice et al., 2012). Similar to the first version of (Ilyas et al., 2017), the updated data set is based on HadCRUT4 ensemble members. For the updated version, the ABC posterior densities of the smoothing parameter and autoregressive weights are determined. These are identified for each of the 2028 months from January 1850 to December 2018. The ABC algorithm (Figure 1) based on the variogram is used to compute the posterior distributions. Uniform priors are considered i.e. $U(0.001, 4)$ and U(1, 4) for the smoothing parameter and autoregressive weight, respectively. The tolerance threshold t is chosen to correspond to the 4% acceptance rate with 250 iterations. It results in 10 sample draws from the posterior densities.

Fitting the multi-resolution lattice kriging model requires a choice of the basis functions and marginal spatial variance. The multi-resolution basis is the same as the one that was chosen for the first version of ensemble data (Ilyas et al., 2017). So a three-level model is chosen such that the number of basis function is greater than the number of spatial locations. The value of $\boldsymbol{\alpha}$ i.e. the marginal spatial variance is estimated as $\boldsymbol{\alpha} = (0.2451, 0.01606, 0.7389)^T$. This is computed over the field with the maximum available information i.e. February 1988. It results from the maximum likelihood estimation as the algorithm in figure 1 is not yet extended to obtain posteriors of $\boldsymbol{\alpha}$ as this parameter has little influence on the uncertainties compared to the others. Other parameters are estimated for each monthly field since the spatial characteristics can vary considerably. The geodesic grid and the great circle distance is used to handle the spherical domain. To implement multi-resolution lattice kriging, the LatticeKrig R package version 6.4 is used. As an example, the posterior densities of smoothing parameter λ and autoregressive weight aw are shown in Figure 2 for one spatial field. These posterior distributions result from the HadCRUT4 spatial field with the minimum spatial coverage i.e. May 1861.

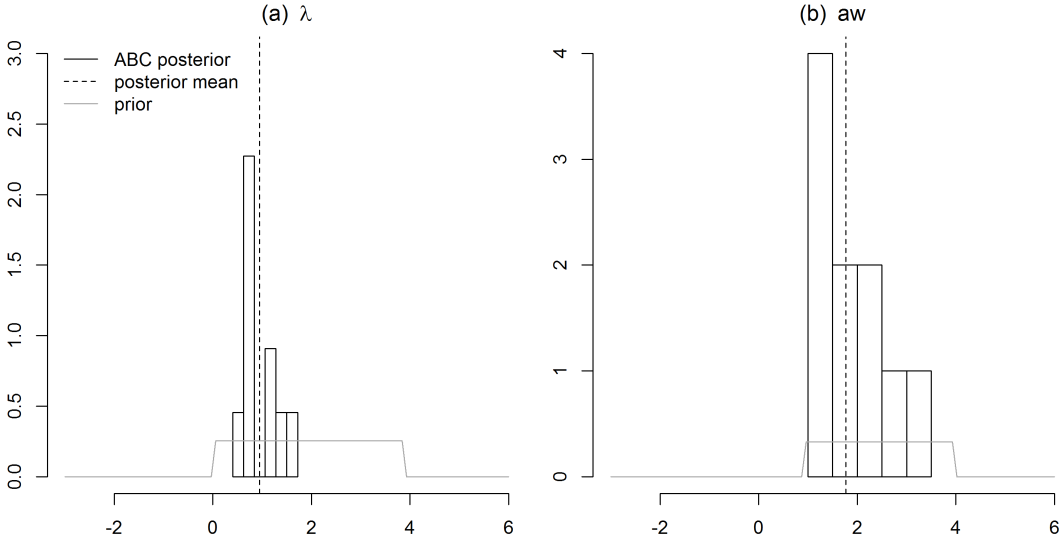

Figure 2. Posterior densities of smoothing parameter λ (a) and autoregressive weight aw (b) for HadCRUT4 spatial field with the minimum spatial coverage i.e. May 1861.

4.2 Spatial field with minimum coverage

ABC-based multi-resolution lattice kriging (Figure 1) is used to predict this sparse spatial field. The spatial predictions and associated uncertainties are shown in Figure 3a and Figure 4a, respectively. The spatial predictions (Figure 3a) are computed using the available spatial observations, multi-resolution basis functions (Section 4.1) and ABC posterior distributions of autoregressive weights aw and smoothing parameter λ (Figure 2). Equation (5) is used for the calculation of the spatial predictor. For comparison with the previous reconstruction (Ilyas et al., 2017) of this spatial field, the spatial predictions based on profile maximum likelihood approach are presented in Figure 3b.

The difference of these reconstructions (Figure 3c) indicate that the spatial predictions using ABC-based multi-resolution lattice kriging (Figure 3a) are higher in magnitude in some regions as compared to the likelihood-based reconstruction (Figure 3b), e.g.for predicted temperatures in the northeast region of North America, southern South America and northern Asia. It can also be observed (Figure 3c) that the ABC based spatial predictions are smaller in certain regions (e.g. Russia and north of Kazakhstan) as compared to the previous reconstruction (Figure 3b). The uncertainties in predictions are shown in Figure 4a that result from ABC-based multi-resolution lattice kriging. These uncertainties in the predictions are computed using equation (6). Figure 4c compares these uncertainties with those resulting from the previous reconstruction (Figure 4b). It can be observed that ABC based multi-resolution lattice kriging is producing higher uncertainty estimates close to the observed spatial sites. This was expected since there is now account of more sources of uncertainty. However, the unobserved grid locations are showing less uncertainties that are resulting from ABC based multi-resolution lattice kriging. This is possibly due to the fact that LKrig function of LatticeKrig R-package undergoes substantial modifications. For the old ensemble (Ilyas et al., 2017), LatticeKrig version 6.2 was used and for hyperparameter ensemble LatticeKrig Version 6.4 was used.

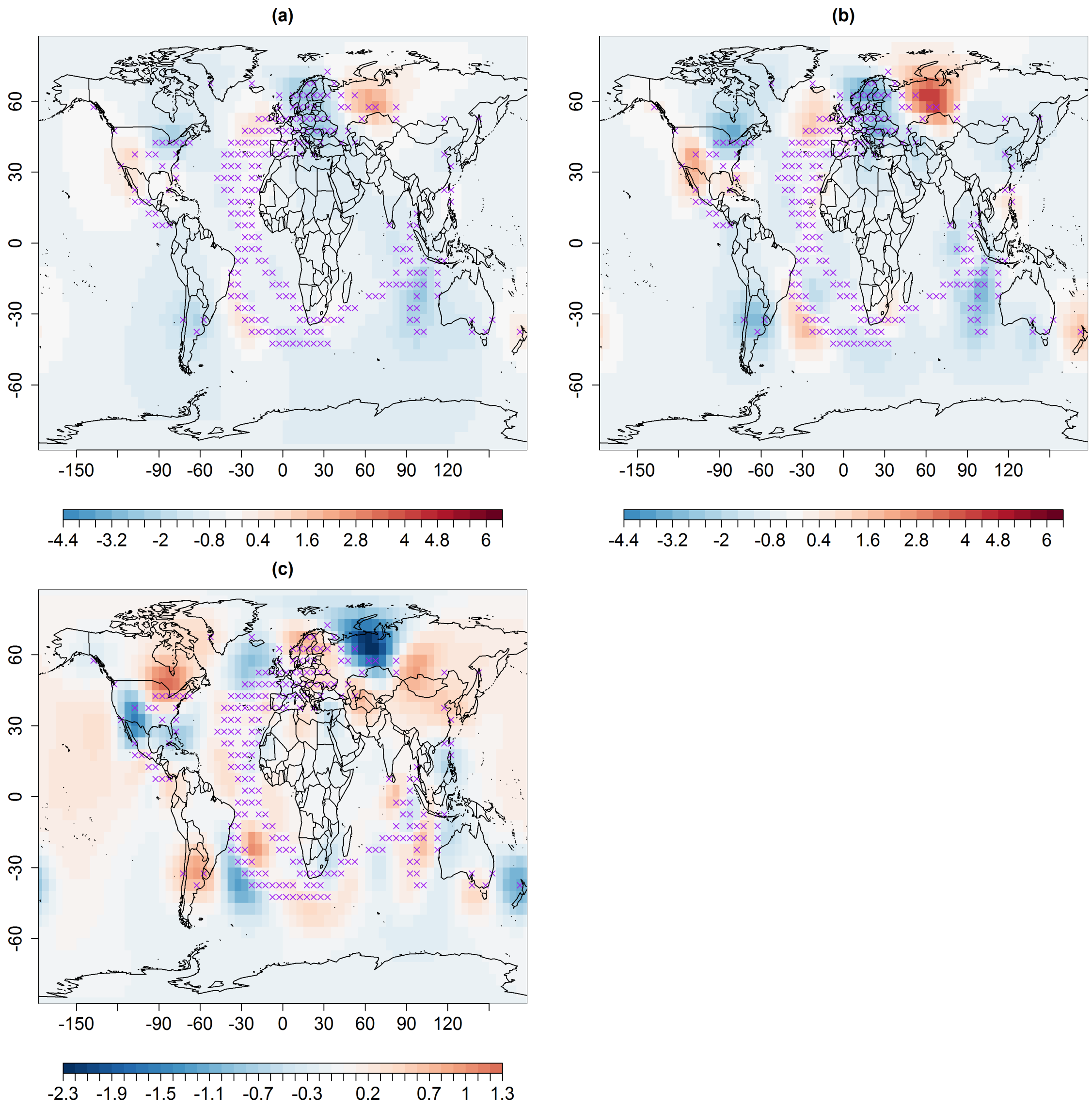

Figure 3. Spatial predictions for May 1861 using multi-resolution lattice kriging based on (a) ABC using variogram (Section 2.1) and (b) profile maximum likelihood approach (Nychka et al., 2015) used to create data in Ilyas et al. (2017). (c) Difference of (a) and (b) i.e. (a)-(b). × signs show observed spatial sites (purple).

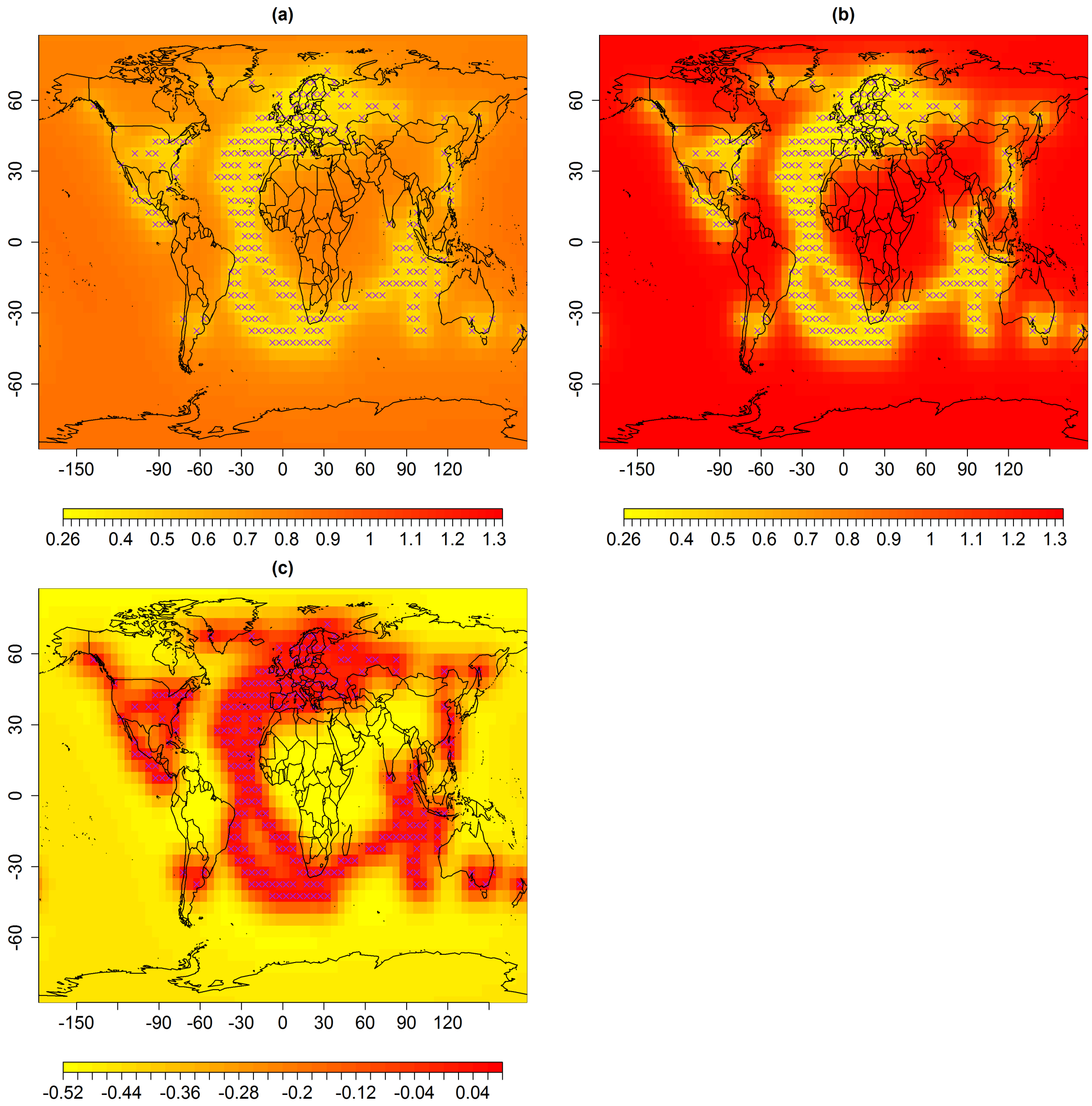

Figure 4. Uncertainties associated with spatial predictions for May 1861 using multi-resolution lattice kriging based on (a) ABC using variogram (Section 2.1) and (b) profile maximum likelihood approach (Nychka et al., 2015) used to create data in Ilyas et al. (2017). (c) Difference of (a) and (b) i.e. (a)-(b). × signs show observed spatial sites (purple).

4.3 100,000 member hyperparameter ensemble

The ABC posterior distributions and model parameters of the multiresolution lattice kriging model (Section 4.1) are used to generate an ensemble. This ensemble data is based on HadCRUT4 temperature data. The HadCRUT4 monthly data set consists of 100 sparse ensemble members. For each of 100 monthly spatial fields of HadCRUT4, a spatially complete 1000 member ensemble is created that samples the coverage and parametric uncertainties of multi-resolution lattice kriging. The resulting 100,000 ensemble members are referred as a hyperparameter temperature ensemble data set. The 1000 members of ensemble generated from each of the 100 HadCRUT4 ensemble members (thus eventually creating $1000 \times 100 = 100,000$ fields) are the random fields drawn from the multivariate conditional normal distribution. These are drawn by conditioning on the HadCRUT4 available field measurements, and sampling the multi-resolution lattice kriging covariance model, namely the variogram based ABC posteriors of autoregressive weights and smoothing parameter. In other words, 100 fields are drawn from the multivariate conditional normal distribution. These are sampled corresponding to each of the 10 draws from the ABC posterior distributions of smoothing parameter and autoregressive weights.

This ensemble data set is generated using High-Performance Computing due to the computational expense. The HadCRUT4 sparse monthly data set spans from 1850 to 2018, 2028 months in total. For each month, the posterior of autoregressive weight aw and smoothing parameter λ are computed using the median ensemble member. Given these posteriors, 1000 coverage samples are drawn for each of 100 HadCRUT4 ensemble members. To achieve sufficiently fast computation, different parts of the data are handled in parallel on different nodes. For this, 100 HadCRUT4 ensemble members are divided into sets of 5. It results in 20 sets each consisting of 5 members. For one time point, the sampling for 100 HadCRUT4 ensemble members is performed in parallel by submitting 20 shared memory parallel jobs with 5 threads. Therefore, a single job that performs computations over 5 ensemble members of a month runs in parallel and takes approximately 66 minutes. The total number of parallel jobs are $20 \times 2028 = 40560$. Therefore, the time required to run these jobs is 40560×66 minutes $= 2676960$ minutes $=$ 61.97 months. Typically, 6 or 7 parallel jobs run simultaneously so it took approximately eight months of wall-clock time to perform these computations.

5 Uncertainties in global mean temperature

The global mean temperature time series are computed for the 100,000 member hyperparameter ensemble data described in Section 4.3. For each ensemble member, the global mean time series is calculated. Figure 5 represents the median time series along with the 95% credible interval. For comparison, Figure 5a also presents the median time series and the uncertainties resulting from an earlier version of the ensemble data (Ilyas et al., 2017), which sampled only combination of observational and coverage uncertainties without uncertianty in the MRLK model.

The impact of including parametric uncertainties can be observed from figure 5 at a global scale. The overall range of annual averages turns out to be wider for hyperparameter ensemble as compared to the ones provided by Ilyas et al. (2017) (Section 4.3) as hyperparameter ensemble spans an added layer of model parametric uncertainties. However, the overall features of the time series resulting from the hyperparameter ensemble and the first version of data (Section 4.3) are mostly similar (Figure 5b). The uncertainties in the temperature anomalies relative to median and HadCRUT4 baseline are higher in the past decades (Figure 5).

https://doi.org/10.5194/amt-2020-454
Preprint. Discussion started: 4 January 2021

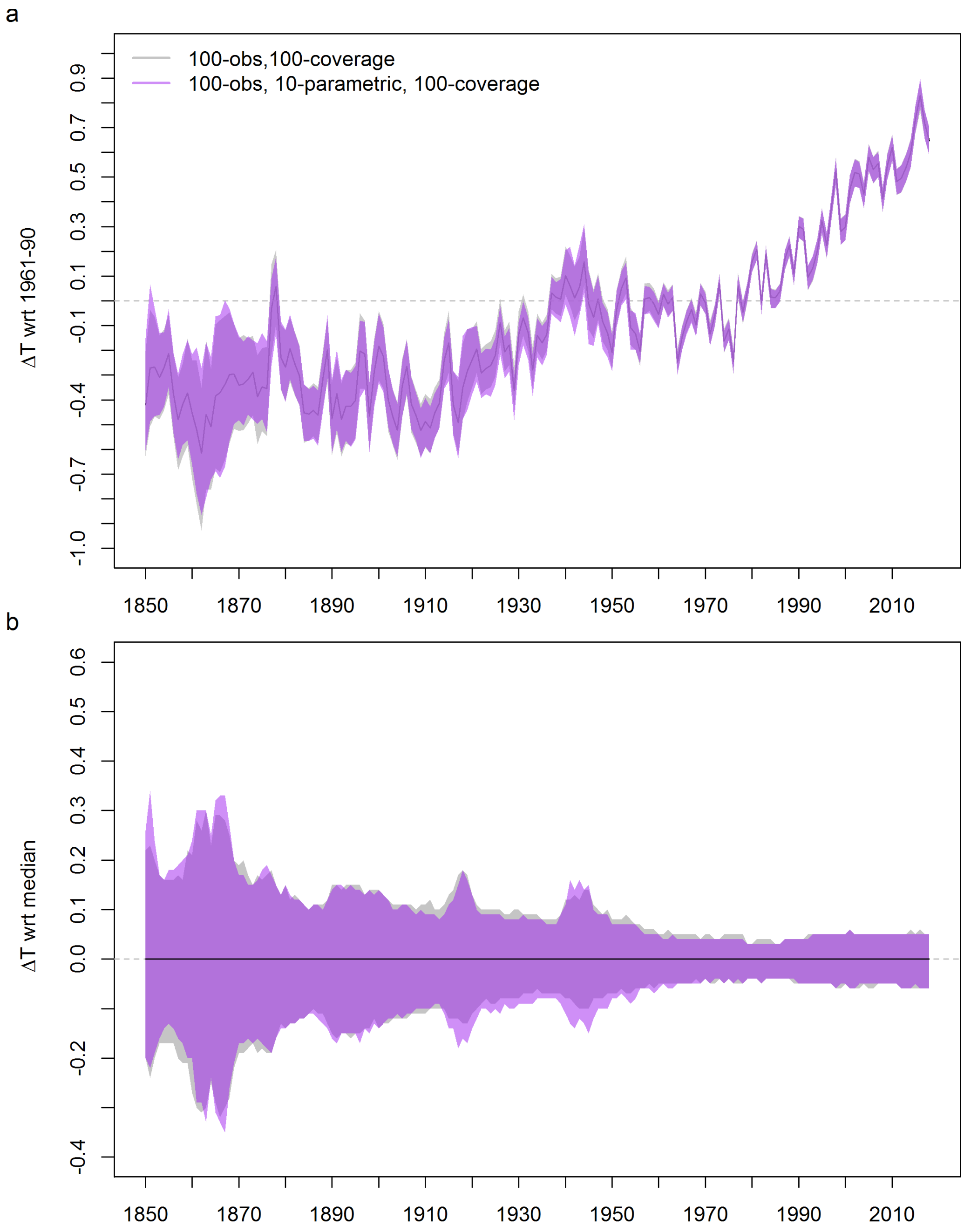

Figure 5. Global mean, annual average temperature anomalies with respect to a) 1961-90 baseline and b) median. 95% credible (purple) and confidence (grey) interval estimates based on the data set created in Ilyas et al. (2017) and hyperparameter ensemble and the data set created in Section 4.3.

https://doi.org/10.5194/amt-2020-454
Preprint. Discussion started: 4 January 2021

Atmospheric Measurement Techniques Discussions

6 Discussion

The hyperparameter ensemble (i.e. an update on Ilyas et al. (2017)) has provided an improved version of the global temperature anomalies since 1850. Instead of classical or frequentist approach, a Bayesian methodology for quantification of uncertainties in large data settings is developed for characterization of uncertainties. The impact of inclusion of parametric uncertainties is evident at a regional (Figure 4) and global scale (Figure 5). Overall, the incorporation of model parametric uncertainties has resulted into larger uncertainties versus the ensemble sampling observational and coverage errors only. The reason of the expansion of these uncertainties is that the uncertainties in the model parameters cascade with observational and coverage uncertainty estimates. However, because of the parametric uncertainties the general understanding of global temperature estimates since 1850 is not significantly changed. The hyperparameter ensemble results in a large ensemble. However, this offsets better assessment of uncertainties in temperature records. Most of the data products address changes in the mean climate whereas this hyperparameter ensemble can be used for studies that aim to explore changing climate variability.

For easy handling of this large data, a subsample of this hyperparameter ensemble is created using Conditioned Latin Hypercube Sampling (CLHS) (Minasny and McBratney, 2006). In practice, Monte Carlo and Latin Hypercube sampling approaches are used to draw samples that approximate the underlying distribution. Usually, a large number of samples are required to achieve good accuracy in traditional Monte Carlo (e.g. Pebesma and Heuvelink, 1999; Olsson and Sandberg, 2002; Olsson et al., 2003; Diermanse et al., 2016). Additionally, the Monte Carlo samples can contain some points clustered closely while other intervals within the space get no sample. On the other hand, the Latin hypercube sampling provides a stratified sampling framework for improved coverage of the k-dimensional input space (e.g. McKay et al., 2000; Helton and Davis, 2003; Iman, 2008; Clifford et al., 2014; Shields and Zhang, 2016; Shang et al., 2020). Conditioned Latin hypercube sampling is an attempt to draw a sample that captures the variation of multiple environmental variables. This sample accurately represents the distribution of the environmental variables over the full range.

To draw a subsample from 100,000 ensemble members, we considered a set of prominent environmental variables i.e. monthly area averages and IPCC AR5 regional means. The conditional Latin hypercube sample is being drawn from the distribution of these environmental variables. The subsample accurately approximates the variation of the set of environmental variables over the full range of these variables. Stating differently, the distributions of the set of environmental variables in the conditioned Latin hypercube sample of size 100 is approximately similar to the distributions of these variables over the full range based on 100,000 ensemble members. As an example, the distribution of a grid box is shown in Figure 6 for May 1861. The full ensemble distribution is based on 100,000 gridboxes (Section-4.3). The subsample distribution results from the conditioned Latin hypercube subsample of 100 gridboxes. Both the distributions overlap mostly. However, the extreme values at the tails are not being captured by the subsample. Also, It is important to note that the subsample ensemble only captures the variation of the specified environmental variables discussed above (i.e. AR5 regional means and monthly area averages). This subsample would not be suitable to be used to explore any other environmental variable. In that case, full hyperparameter ensemble is suggested to be used.

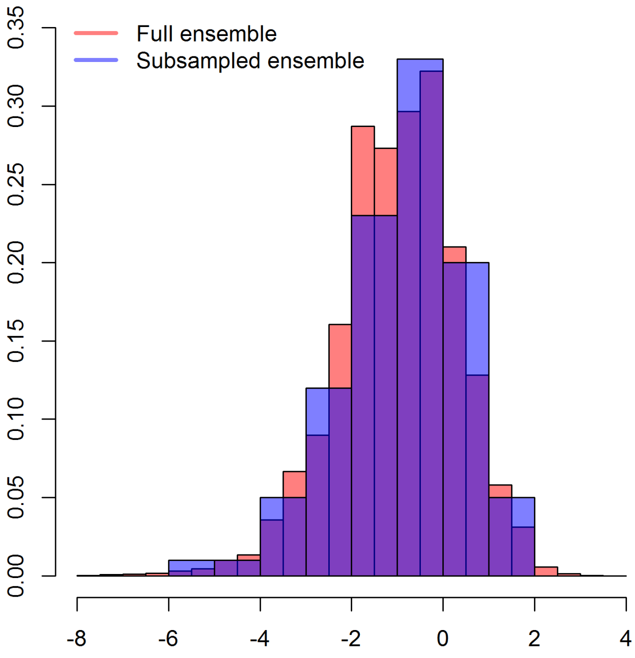

Figure 6. Distribution of the grid box, (lon, lat) = (72.5°, 32.5°) that includes Lahore for May-1861. Full ensemble consists of 100,000 gridboxes (Section- 4.3) and subsampled ensemble consists of 100 gridboxes based on Latin Hypercube sampling.

Acknowledgements. This research was supported by a UCL Overseas Research Scholarship (M.I.), a University of the Punjab Overseas Scholarship (M.I.) and Research Grant (M.I), and National Center for Atmospheric Research (NCAR) (M.I.). This research would not have been possible without the assistance and generosity of the Met. Office's Hadley Centre and the University of East Anglia's Climate Research Unit. The authors acknowledge the use of the UCL Legion High Performance Computing Facility (Legion@UCL), and associated support services, in the completion of this work.

References

M.A. Beaumont. Approximate Bayesian computation in evolution and ecology. *Annual Review of Ecology, Evolution, and Systematics*, 41: 379–406, 2010.

M.A. Beaumont. Approximate bayesian computation. *Annual Review of Statistics and its Application*, 6:379–403, 2019.

M.A. Beaumont, W. Zhang, and D.J. Balding. Approximate Bayesian computation in population genetics. *Genetics*, 162(4):2025–2035, 2002.

P. Brohan, J.J. Kennedy, I. Harris, S.F.B. Tett, and P.D. Jones. Uncertainty estimates in regional and global observed temperature changes: A new data set from 1850. *Journal of Geophysical Research: Atmospheres*, 111(D12), 2006.

A.G. Busetto and J.M. Buhmann. Stable Bayesian parameter estimation for biological dynamical systems. In *Computational Science and Engineering, 2009. CSE'09. International Conference on*, volume 1, pages 148–157. IEEE, 2009.

L. Cao, Z. Yan, P. Zhao, Y. Zhu, Y. Yu, G. Tang, and P. Jones. Climatic warming in China during 1901–2015 based on an extended dataset of instrumental temperature records. *Environmental Research Letters*, 12(6):064005, 2017. URL http://stacks.iop.org/1748-9326/12/i=6/a=064005.

D. Clifford, J.E. Payne, M.J. Pringle, R. Searle, and N. Butler. Pragmatic soil survey design using flexible Latin hypercube sampling. *Computers & Geosciences*, 67:62–68, 2014.

K.W. Cowtan and G. Robert. Coverage bias in the HadCRUT4 temperature series and its impact on recent temperature trends. *Quarterly Journal of the Royal Meteorological Society*, 140(683):1935–1944, 2014.

F.L.M. Diermanse, D.G. Carroll, J.V.L. Beckers, and R. Ayre. An efficient sampling method for fast and accurate Monte Carlo Simulations. *Australasian Journal of Water Resources*, 20(2):160–168, 2016.

P. Domonkos and J. Coll. Homogenisation of temperature and precipitation time series with ACMANT3: Method description and efficiency tests. *International Journal of Climatology*, 37(4):1910–1921, 2017.

R.J.H. Dunn, K.M. Willett, C.P. Morice, and D.E. Parker. Pairwise homogeneity assessment of HadISD. *Climate of the Past*, 10(4):1501, 2014.

R.J.H. Dunn, K.M. Willett, D.E. Parker, and L. Mitchell. Expanding HadISD: Quality-controlled, sub-daily station data from 1931. *Geoscientific Instrumentation, Methods and Data Systems*, 5(2):473, 2016.

R. Dutta, B. Chopard, J. Lätt, F. Dubois, K.Z. Boudjeltia, and A. Mira. Parameter estimation of platelets deposition: Approximate Bayesian computation with high performance computing. *arXiv preprint arXiv:1710.01054*, 2017.

J.M. Edwards, J.R. McGregor, M.R. Bush, and F.J. Bornemann. Assessment of numerical weather forecasts against observations from Cardington: seasonal diurnal cycles of screen-level and surface temperatures and surface fluxes. *Quarterly Journal of the Royal Meteorological Society*, 137(656):656–672, 2011.

H. Feng, X. Zhao, F. Chen, and L. Wu. Using land use change trajectories to quantify the effects of urbanization on urban heat island. *Advances in Space Research*, 53(3):463–473, 2014.

N. Glanemann, S.N. Willner, and A. Levermann. Paris Climate Agreement passes the cost-benefit test. *Nature Communications*, 11(1):1–11, 2020.

Elizabeth Jane Good. An in situ-based analysis of the relationship between land surface "skin" and screen-level air temperatures. *Journal of Geophysical Research: Atmospheres*, 121(15):8801–8819, 2016.

J.P. Gosling, S.M. Krishnan, G. Lythe, B. Chain, C. MacKay, and C. Molina-París. A mathematical study of CD8+ T cell responses calibrated with human data. *arXiv preprint arXiv:1802.05094*, 2018.

J. Hansen, R. Ruedy, M. Sato, and K. Lo. Global surface temperature change. *Reviews of Geophysics*, 48(4), 2010.

Z. Hausfather, K. Cowtan, M.J. Menne, and C.N. Williams. Evaluating the impact of us historical climatology network homogenization using the us climate reference network. *Geophysical Research Letters*, 43(4):1695–1701, 2016.

J.C. Helton and F.J. Davis. Latin hypercube sampling and the propagation of uncertainty in analyses of complex systems. *Reliability Engineering & System Safety*, 81(1):23–69, 2003.

Boyin Huang, Matthew J Menne, Tim Boyer, Eric Freeman, Byron E Gleason, Jay H Lawrimore, Chunying Liu, J Jared Rennie, Carl J Schreck III, Fengying Sun, et al. Uncertainty estimates for sea surface temperature and land surface air temperature in NOAAGlobalTemp version 5. *Journal of Climate*, 33(4):1351–1379, 2020.

M. Ilyas, C.M. Brierley, and S. Guillas. Uncertainty in regional temperatures inferred from sparse global observations: Application to a probabilistic classification of El Niño. *Geophysical Research Letters*, 44(17):9068–9074, 2017.

R.L. Iman. *Latin hypercube sampling*. Wiley Online Library, 2008.

M. Ishii, A. Shouji, S. Sugimoto, and T. Matsumoto. Objective analyses of sea-surface temperature and marine meteorological variables for the 20th century using ICOADS and the Kobe collection. *International Journal of Climatology*, 25(7):865–879, 2005.

P. Jones. The reliability of global and hemispheric surface temperature records. *Advances in Atmospheric Sciences*, 33(3):269–282, 2016.

P.D. Jones, D.H. Lister, T.J. Osborn, C. Harpham, M. Salmon, and C.P. Morice. Hemispheric and large-scale land-surface air temperature variations: An extensive revision and an update to 2010. *Journal of Geophysical Research: Atmospheres*, 117(D5), 2012.

J.J Kennedy. A review of uncertainty in in situ measurements and data sets of sea surface temperature. *Reviews of Geophysics*, 52(1):1–32, 2014.

J.J. Kennedy, N.A. Rayner, R.O. Smith, D.E. Parker, and M. Saunby. Reassessing biases and other uncertainties in sea surface temperature observations measured in situ since 1850: 2. Biases and homogenization. *Journal of Geophysical Research: Atmospheres*, 116(D14), 2011a.

J.J. Kennedy, N.A. Rayner, R.O. Smith, D.E. Parker, and M. Saunby. Reassessing biases and other uncertainties in sea surface temperature observations measured in situ since 1850: 1. Measurement and sampling uncertainties. *Journal of Geophysical Research: Atmospheres*, 116(D14), 2011b.

N. Lenssen, G. Schmidt, J. Hansen, M. Menne, A. Persin, R. Ruedy, and D. Zyss. Improvements in the GISTEMP uncertainty model. *Journal of Geophysical Research: Atmospheres*, 124(12):6307–6326, 2019. https://doi.org/10.1029/2018JD029522.

M.D. McKay, R.J. Beckman, and W.J. Conover. A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics*, 42(1):55–61, 2000.

K.A. McKinnon, A. Poppick, E. Dunn-Sigouin, and C. Deser. An “Observational Large Ensemble” to compare observed and modeled temperature trend uncertainty due to internal variability. *Journal of Climate*, 30(19):7585–7598, 2017.

M.J. Menne, C.N. Williams, B.E. Gleason, J.J. Rennie, and J.H. Lawrimore. The global historical climatology network monthly temperature dataset, version 4. *Journal of Climate*, 31(24):9835–9854, 2018.

S.F. Milton and P. Earnshaw. Evaluation of surface water and energy cycles in the Met Office global NWP model using CEOP data. *Journal of the Meteorological Society of Japan. Ser. II*, 85:43–72, 2007.

Budiman Minasny and Alex B McBratney. A conditioned latin hypercube method for sampling in the presence of ancillary information. *Computers & geosciences*, 32(9):1378–1388, 2006.

A. Moberg, H. Alexandersson, H. Bergström, and P.D. Jones. Were southern Swedish summer temperatures before 1860 as warm as measured? *International Journal of Climatology*, 23(12):1495–1521, 2003.

C.P. Morice, J.J. Kennedy, N.A. Rayner, and P.D. Jones. Quantifying uncertainties in global and regional temperature change using an ensemble of observational estimates: The HadCRUT4 data set. *Journal of Geophysical Research: Atmospheres (1984–2012)*, 117(D8), 2012.

C.P. Morice, J.J. Kennedy, N.A. Rayner, J.P. Winn, E. Hogan, R.E. Killick, R.J.H. Dunn, T.J. Osborn, P.D. Jones, and I.R. Simpson. An updated assessment of near-surface temperature change from 1850: the HadCRUT5 dataset. *Journal of Geophysical Research: Atmospheres*, page e2019JD032361, 2020.

D. Nychka, S. Bandyopadhyay, D. Hammerling, F. Lindgren, and S. Sain. A multi-resolution Gaussian process model for the analysis of large spatial data sets. *Journal of Computational and Graphical Statistics*, 24(2):579–599, 2015.

A. Olsson, G. Sandberg, and O. Dahlblom. On Latin hypercube sampling for structural reliability analysis. *Structural safety*, 25(1):47–68, 2003.

A.M.J. Olsson and G.E. Sandberg. Latin hypercube sampling for stochastic finite element analysis. *Journal of Engineering Mechanics*, 128 (1):121–125, 2002.

D.E. Parker. Effects of changing exposure of thermometers at land stations. *International Journal of Climatology*, 14(1):1–31, 1994.

D.E. Parker. Urban heat island effects on estimates of observed climate change. *Wiley Interdisciplinary Reviews: Climate Change*, 1(1): 123–133, 2010.

E.J. Pebesma and G.B.M. Heuvelink. Latin hypercube sampling of Gaussian random fields. *Technometrics*, 41(4):303–312, 1999.

Jonathan K Pritchard, Mark T Seielstad, Anna Perez-Lezaun, and Marcus W Feldman. Population growth of human y chromosomes: a study of y chromosome microsatellites. *Molecular biology and evolution*, 16(12):1791–1798, 1999.

N.A. Rayner, P. Brohan, D.E. Parker, C.K. Folland, J.J. Kennedy, M. Vanicek, T.J. Ansell, and S.F.B. Tett. Improved analyses of changes and uncertainties in sea surface temperature measured in situ since the mid-nineteenth century: The HadSST2 dataset. *Journal of Climate*, 19 (3):446–469, 2006.

R. Rhode, R.A. Muller, R. Jacobsen, E. Muller, S. Perlmutter, A. Rosenfeld, J. Wurtele, D. Groom, and C. Wickham. A new estimate of the average earth surface land temperature spanning 1753 to 2011. *Geoinformatics & Geostatistics: An Overview*, 2013.

R. Rohde. Comparison of Berkeley Earth, NASA GISS, and Hadley CRU averaging techniques on ideal synthetic data. *Berkeley Earth Memo, January*, 2:013, 2013.

X. Shang, T. Chao, P. Ma, and M. Yang. An efficient local search-based genetic algorithm for constructing optimal latin hypercube design. *Engineering Optimization*, 52(2):271–287, 2020.

M.D. Shields and J. Zhang. The generalization of Latin hypercube sampling. *Reliability Engineering & System Safety*, 148:96–108, 2016.

T.M. Smith, R.W. Reynolds, T.C. Peterson, and J. Lawrimore. Improvements to NOAA's historical merged land-ocean surface temperature analysis (1880-2006). *Journal of Climate*, 21(10):2283–2296, 2008.

B. Trewin. Exposure, instrumentation, and observing practice effects on land temperature measurements. *Wiley Interdisciplinary Reviews: Climate Change*, 1(4):490–506, 2010.

R.S. Vose, D. Arndt, V.F. Banzon, D.R. Easterling, B. Gleason, B. Huang, E. Kearns, Jay H Lawrimore, Matthew J Menne, Thomas C Peterson, et al. NOAA's merged land–ocean surface temperature analysis. *Bulletin of the American Meteorological Society*, 93(11): 1677–1685, 2012.

S.D. Woodruff, S.J. Worley, S.J. Lubker, Z. Ji, J. Eric F., D.I. Berry, P. Brohan, E.C. Kent, R.W. Reynolds, S.R. Smith, et al. ICOADS Release 2.5: extensions and enhancements to the surface marine meteorological archive. *International journal of climatology*, 31(7): 951–967, 2011.