Norsk Regnesentral
NORWEGIAN COMPUTING CENTER

Note

A combined estimate of global temperature

Note no SAMBA/14/21
Authors Peter Guttorp
Peter F. Craigmile
Date April 30, 2021

Norsk Regnesentral
NORWEGIAN COMPUTING CENTER

The authors

Peter Guttorp is Professor II at the Norwegian Computing Center. Peter F. Craigmile is Professor of Statistics at the Ohio State University, USA.

Norwegian Computing Center

Norsk Regnesentral (Norwegian Computing Center, NR) is a private, independent, non-profit foundation established in 1952. NR carries out contract research and development projects in information and communication technology and applied statistical-mathematical modelling. The clients include a broad range of industrial, commercial and public service organisations in the national as well as the international market. Our scientific and technical capabilities are further developed in co-operation with The Research Council of Norway and key customers. The results of our projects may take the form of reports, software, prototypes, and short courses. A proof of the confidence and appreciation our clients have in us is given by the fact that most of our new contracts are signed with previous customers.

Norsk Regnesentral
Norwegian Computing Center
Postboks 114, Blindern
NO-0314 Oslo, Norway

Besøksadresse
Office address
Gaustadalléen 23
NO-0373 Oslo, Norway

Telefon · telephone
(+47) 22 85 25 00
Telefaks · telefax
(+47) 22 69 76 60

Internett · internet
www.nr.no
E-post · e-mail
nr@nr.no

Title	**A combined estimate of global temperature**
Authors	**Peter Guttorp , Peter F. Craigmile**
Date	April 30, 2021
Publication number	SAMBA/14/21

Abstract

Recently several global temperature series have been updated using new data sets, new methods, and for a statistician most importantly, assessments of their uncertainties. This enables us to produce a timely estimate of the annual global mean temperature with a combined estimate of uncertainty. We describe the hierarchical model we propose, and a Bayesian scheme for fitting the model. In addition, we test the sensitivity to the results to each of the series, identifying groups of data products that act similar to one another. Using the combined estimate of the global temperature series, we estimate that the probability that 2020 was the warmest year on record is 0.41.

Keywords	Globalwarming; latent model; combination of data sources; sensitivity analysis
Target group	Statisticians, Climate scientists
Availability	Open
Project	Climate Futures
Project number	220993
Research field	Statistics
Number of pages	20

1 Introduction

Global mean temperature, while not a sensitive indicator of climatic change (Hegerl et al., 2006), is a commonly used diagnostic of said change. In order to estimate this quantity from data, mainly land-based temperature measurements, ship-based sea surface temperatures, bouys, and floats, different groups use different approaches, and to some extent different raw data. There are one private and four governmental groups that provide estimates of global monthly or annual mean temperatures. The series are subject to uncertainty, and different groups have different approaches to assess this as well. For a statistician, it seems reasonable to try to combine the series into a single estimate. This ought to reduce the uncertainty in the estimate. Our approach is to assume that each of the series measure (with series-dependent errors) the same quantity, the actual global mean temperature, and try to use the data and their uncertainties to estimate this quantity and determine the uncertainty of that estimate. We have chosen recent updates of all five global annual mean temperature series that come with estimates of uncertainty. Recent updates to the methods in several of these series make it timely to publish this analysis. While we will try to explain how the uncertainties have been calculated, we will not attempt to make any judgment as to which approach is better.

In Section 2 we first describe the current global mean temperature series and some of the statistical issues surrounding them, and in Section 3 we explain our methodology. In Section 4 we provide the results, summarizing our posterior estimate of the combined posterior global temperature anomalies, as well as characteristics of the trend and time series dependence of this anomaly process. We explore the sensitivity of our results to the omission of each of the five global anomaly series that we include in our analysis, and investigate the probability that 2020 was the warmest year on record. We close with a discussion in Section 5.

2 Current methods

The Global Historical Climatology Network (GHCN) is a collection of land based weather stations, maintained by the National Oceanic and Atmospheric Administration (NOAA) in the United States. The largest set of stations is the monthly set, which in its current version 4 has about 26 000 stations with some records going back to the 18th century (Menne et al., 2018). The data sets have been homogenized to take account of instrument changes and station moves. Recently, many additional stations have been added using the International Surface Temperature Initiative (ISTI; Rennie et al. (2014)). In particular, this has improved the coverage of high northern latitudes.

The International Comprehensive Ocean-Atmosphere Data Set (ICOADS), also maintained by NOAA, contains records of marine surface data, some going back to the 18th century (Freeman et al., 2017). The older data are nearly all based on commercial or military ship logs. The data set has roughly half a billion observations. Bias-corrected analy-

ses are included in ERSSTv5 (Huang et al., 2017) and HadSST.4 (Kennedy et al., 2019).

The UK Hadley Center and University of East Anglia Climate Research Unit (HadCRU) estimate global mean land temperature by first dividing each hemisphere into 5 degree by 5 degree grids, within each of which all selected GHCN monitoring stations are averaged (Morice et al., 2020). Stations with suspiciously large deviations from its neighbours are eliminated. The sea surface temperature is similarly calculated using ICOADS data. The global mean temperature is then calculated by averaging water and land grid squares with weights corresponding to the proportion of the grid square that is land and ocean, respectively, as well as the area of the grid square compared to the area of Earth. Grid squares without observations (particularly in the Arctic) were previously ignored, but are now estimated using Gaussian spatial estimation.

The US National Aeronautical and Space Administration Goddard Institute of Space Studies (GISS) use a finer grid with 8 000 equal area grid cells, and compute the grid cell averages by distance weighted averaging (Hansen et al., 2010; Lenssen et al., 2019). The land data are from GHCNv4, and the ocean data are from ERSSTv4. GISS also uses some satellite data to correct for urbanicity.

The Japanese Meteorological Agency (JMA) uses GHCN data through 2000, and WMO CLIMAT reports from 2001 on. They use their own sea surface analysis COBE-SST (Ishii et al., 2005), which is partly based on ICOADS data. The calculation of global mean temperature is similar to previous ones of the Hadley Center (leaving out areas of no observations).

NOAA calculates an empirical Karhunen-Loeve expansion based on the observed spatial covariance during a reference period, and estimates the average temperature in 5 degree by 5 degree grid squares without observations from weighted averages of station values based on the expected spatial correlation between the station and the grid square center (Huang et al., 2019; Vose et al., 2012; Zhang et al., 2019). The land data are from GHCNv4, and the ocean data are from ERSSTv5.

A final global temperature series is produced by the Berkeley Earth group. They use a 1 degree by 1 degree grid for land data (Rohde and Hausfather, 2020), and use tools of spatial statistics to estimate the temperature in land areas without observations In addition they do not homogenize data. Rather, when a station is moved or is known to have changed instrumentation, it is considered a new station, independent of the old one. Their land data has about 1.6 million observations at over 40 000 stations, and are obtained by combining 16 data sources. The ocean data uses an interpolated version of the Hadley Center product HadSST3.

The different series all use monthly data (many stations overlap between the data sets) which are standardized by subtracting monthly averages for a baseline interval of 20–30 years, resulting in anomalies. Different groups calculate the anomalies differently. Rohde et al. (2013) contains more detailed descriptions of the calculation of some of the series.

For our analysis, we subset each data product to the years 1880–2020, and generate tem-

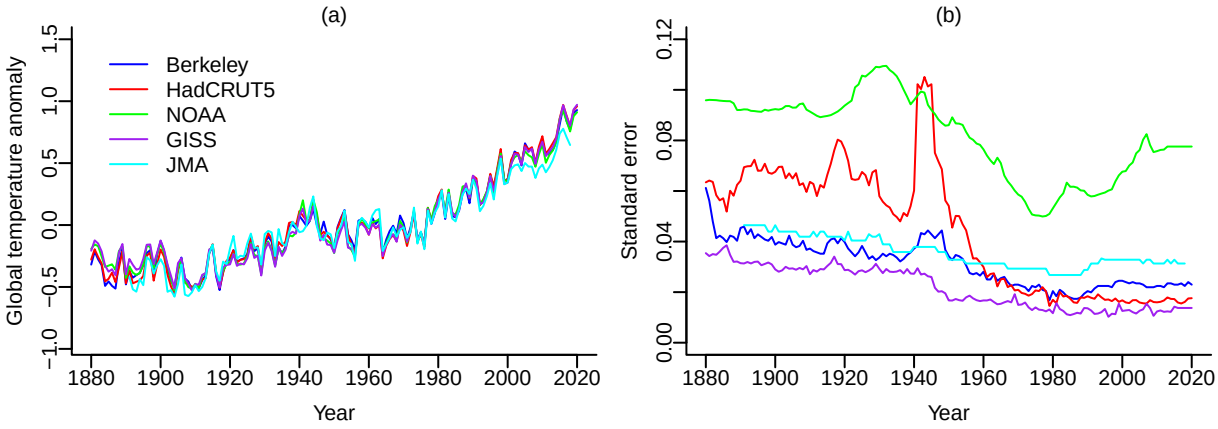

Figure 1. (a) Time series plots of the global temperature anomalies for the five data products; (b) Plots of the standard errors, by year.

perature anomalies by subtracting the mean for the entire period for each series. However, for the JMA series, there were no data available for the years 1880–1890 and 2019–2020, and our subtraction of the mean took account of this shortened time period.

Figure 1(a) shows a time series plot of these global anomalies. There are strong positive pairwise correlations in range of 0.967–0.996, driven by the trend. While there are many commonalities among the series (e.g., all the series are warmer in 1940–1950 and from 1980 onwards), there are also year-by-year differences. For example, the series tend to disagree on the values of the global temperature anomaly before 1900, and the JMA series tends to be cooler in the later years.

2.1 Statistical difficulties

In order to do proper spatial estimation of the global temperature field, it is necessary to model the spatial covariance function. Since the covariance function has to be defined on a sphere, it is necessary to use a theoretical one (possibly with estimated parameters) rather than an empirical one (see Gneiting (2013), for a discussion).

The creation of anomalies from data (or from global reconstructions) for a given reference point has two purposes: it reduces the spatial correlation, and it serves to remove a seasonal effect. For a statistician this would be estimated using the entire data set, while for climatologists this is done using a reference period. As a consequence, the temporal structure of the anomalies during the reference period is less pronounced than for other periods. In addition, if the reference period contains an unusual seasonal pattern, the rest of the anomalies will contain residual seasonal variation. In some applications this could influence the results.

When approximating the Earth by a sphere, the concept of stationarity (translation invariance) translates to isotropy (rotation invariance). Most of the above analyses use isotropic structures. It does not appear likely that most climate processes are isotropic. A proper nonstationary estimation of global mean temperature, but for computational reasons based on a small subset of the data available, can be found in the PhD disserta-

tion of Barnali Das at the University of Washington (Das, 2000).

2.2 Estimates of uncertainty

The five series (Berkeley Earth, Hadley, NOAA, GISS, and JMA) calculate uncertainties of each global estimate. Berkeley Earth computes standard geostatistical estimate of uncertainty. The Hadley analysis computes standard deviations of the analysis (spatial estimation) error, which includes measurement error, sampling error and SST bias correction error, and coverage error over parts of the globe considered too far from observations to perform spatial estimation (those areas where the ratio of posterior to prior variance is higher than 25%). The NOAA series has standard errors corresponding to coverage error and SST bias correction error. The GISS uncertainty corresponds to land station homogenization errors, ocean temperature bias correction error, spatial interpolation and coverage error, as well as parametric uncertainty. Finally, JMA uses a similar approach to Hadley to assess uncertainty (although no uncertainty is reported after 2016 – in our analysis we assume the uncertainty for 2017 and 2018 is the same as the uncertainty reported for 2016).

Figure 1(b) shows time series plot of these standard errors. We notice that the NOAA series uncertainty generally is higher than the other series' uncertainties. Around World War 2 the Hadley series expresses high uncertainty, presumably due to higher observation bias in ocean data collected mainly from naval ship records.

3 Methods summary

In this section we present a hierarchical statistical model that we employ to produce a combined estimate of the global temperature anomaly using the different global data products. We also discuss the different models that we fit to the different data sources available to us.

3.1 A model to reconstruct global temperature anomalies

In our model we assume a common time scale indexed by $t = 1, \ldots, N$, but allow the data products to be missing at certain time indexes. This also allows us to understand the effect of removing certain data products from the model, but also accounts for the JMA series that has missing data relative to the other data products.

Suppose that we have J data products, and let $j = 1, \ldots, J$ index the data products. For data product j, let $\{D_{t,j} : t = 1, \ldots, N\}$ denote the global mean temperature anomaly series and let $\{v_{j,t} : t = 1, \ldots, N\}$ denote the time-varying measure of variability that accompanies each global mean anomaly series. To allows for the possibility of missing values, we set $o^D_{j,t} = 1$ if we observe an anomaly at time index t for data product j and $o^D_{j,t} = 0$ otherwise.

Now let $\{Y_t : t = 1, \ldots, N\}$ denote the latent global temperature anomaly that we wish

to infer upon. We assume that

$$Y_t \quad = \quad \mu_t + \nu_t, \quad t = 1, \ldots, N.$$

Here the trend component $\{\mu_t\}$ is modeled by a linear combination of b basis functions $\{\boldsymbol{x}^T(t)\}$:

$$\mu_t \quad = \quad \boldsymbol{x}^T(t)\,\boldsymbol{\beta}, \quad t = 1, \ldots, N.$$

The serial dependence about the trend, $\{\nu_t\}$, is modeled using a stationary Gaussian autoregressive process of order p (AR(p)). To ensure that the AR process $\{\nu_t\}$ is stationary, we parameterize the model using the partial autocorrelation function (PACF), which uniquely defines the process. Since the PACF is zero at lags $h > p$ for an AR(p) process, let $\boldsymbol{\psi}$ denote the p-vector of the first p lags of the PACF. Also, since the PACF at each lag $h = 1, \ldots, p$ is constrained to lie between -1 and 1, we choose to transform the PACF to the real line using the link function $g(x) = \log((1+x)/(1-x))$. Applying the link function elementwise, we define $\boldsymbol{\eta} = g(\boldsymbol{\psi})$ to be the transformed PACF vector. Given $\boldsymbol{\eta}$ or $\boldsymbol{\psi}$, and an innovation variance $\sigma^2 > 0$, we can calculate the covariance of the AR(p) process using the Levinson-Durbin recursions (e.g. Brockwell and Davis, 2002).

When we observe an observation for data product $j = 1, \ldots, J$ and time index t (i.e., when $o^D_{t,j} = 1$), then

$$D_{j,t} = Y_t + \delta_{j,t} + \epsilon_{j,t}, \quad t = 1, \ldots, T. \tag{1}$$

This model (1) assumes that the global temperature anomaly for each data product is equal to the latent global temperature, plus a discrepancy term $\{\delta_{j,t}\}$ that captures the variability between the data products, as well as a term that captures the natural variability of each data product $\{\epsilon_{j,t}\}$. Using the measure of variability, $\{v_{j,t} : t = 1, \ldots, N\}$, that comes with each data product j, we assume that $\{\epsilon_{j,t} : t = 1, \ldots, N\}$ is an independent Gaussian process with mean zero and variance $\text{var}(\epsilon_{j,t}) = v_{j,t}$. We assume that $\{\delta_{j,t} : t = 1, \ldots, T\}$ is a set of independent N(0, τ^2) random variables for each j, and assume independence of these discrepancy terms over the data products.

3.2 Fitting the Bayesian model

Using a Bayesian approach, we will assume prior distributions for the hyperparameters τ^2, $\boldsymbol{\beta}^T$, $\boldsymbol{\eta}^T$, σ^2, that appear in our model. We first assume mutual independence between each parameter. Then for the variance τ^2 of the data product discrepancies we assume an inverse gamma distribution with shape parameter 0.01 and rate 0.01. For the trend coefficients $\boldsymbol{\beta}$, we assume a b-variate normal distribution with mean $\mathbf{0}_b$ and covariance $10\boldsymbol{I}_b$. For the AR(p) process $\{\nu_t\}$, we assume that the transformed PACF parameters $\boldsymbol{\eta}$ follow a p-variate normal distribution with mean $\mathbf{0}_p$ and covariance $0.5\boldsymbol{I}_p$ and the innovation variance σ^2 follows an inverse gamma distribution with shape parameter 0.01 and rate 0.01.

Including the latent global temperature process observed at all time indexes, $\boldsymbol{Y} = (Y_1, \ldots, Y_N)^T$, our complete vector of parameters in our hierarchical model is

$$\boldsymbol{\theta} \quad = \quad (\boldsymbol{Y}, \tau^2, \boldsymbol{\beta}^T, \boldsymbol{\eta}^T, \sigma^2)^T.$$

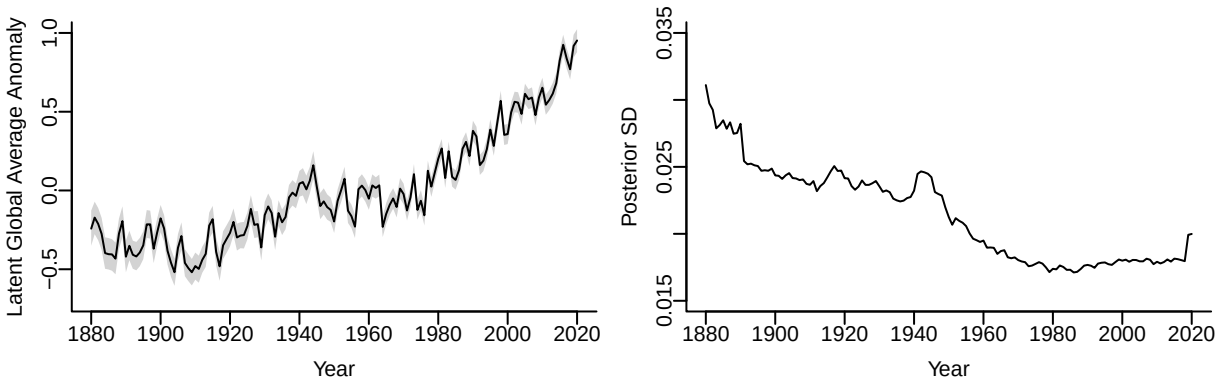

Figure 2. (a) Time series plots of the posterior mean global temperature anomalies (black lines), along with simultaneous 95% credible intervals (gray regions). (b) A plot of the posterior standard deviation of the global temperature anomalies by year.

Then the posterior distribution of the parameters given the data products is

$$\pi(\boldsymbol{\theta}|\{\boldsymbol{D}_j\}) \quad \propto \quad \left[\prod_{j=1}^{J} f(\boldsymbol{D}_j|\boldsymbol{Y},\tau^2)\right] \pi(\boldsymbol{Y}|\boldsymbol{\beta},\boldsymbol{\eta},\sigma^2)\ \pi(\tau^2)\ \pi(\boldsymbol{\beta})\ \pi(\boldsymbol{\eta})\ \pi(\sigma^2).$$

This posterior distribution is not available in closed form, and so we used a Markov chain Monte Carlo (MCMC) algorithm to sample from the parameters given the data products and climate model runs. Details of the MCMC algorithm and how we sample the model discrepancy terms are given in Appendix A.

In every model we use $b = 5$ b-spline basis functions to capture long term trends on a scale of approximately 25 years in the latent climate anomaly. We also varied the order p of the AR(p) model. There was little appreciable differences between the results with $p = 2$ and $p = 4$, and so we will only discuss the AR(4) model results. (Also, see Craigmile and Guttorp (2019) for a time series analysis of the Berkeley Earth series that suggested that an AR(4) model for the errors was the best fit.)

4 Results

We fit the hierarchical statistical model to the five global anomaly series along with their estimates of uncertainty, using MCMC with Gibbs Sampling. We ran the chain for 500 000 iterations after burn-in and thinned the chains by keeping every 10th sample. We checked convergence of the chains using trace plots, running multiple chains from different starting values.

4.1 Posterior global temperature anomalies

Figure 2(a) shows a plot of the posterior mean global temperature anomalies by year calculated using our hierarchical model fit to the five global anomaly series along with their estimates of uncertainty (black lines). The gray regions show simultaneous 95% cred-

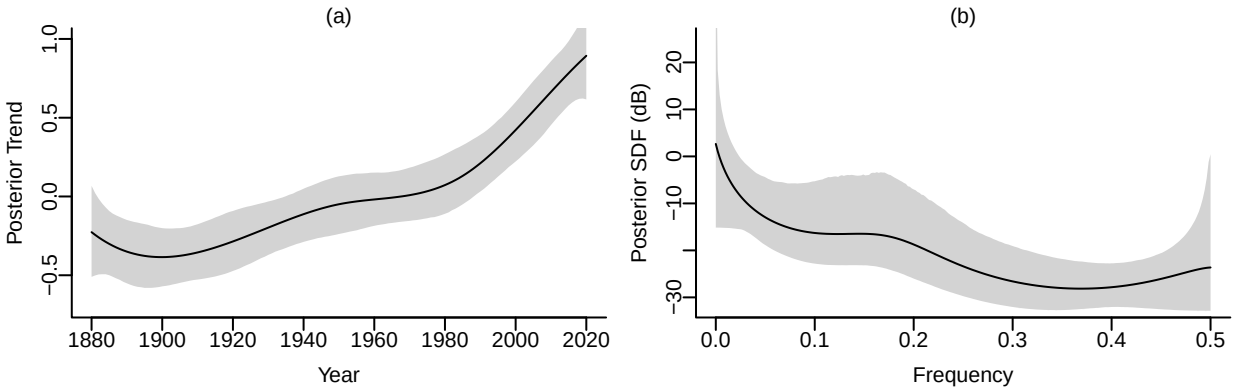

Figure 3. (a) Time series plots of the posterior mean trend (black lines), along with simultaneous 95% credible intervals (gray regions); (b) Posterior mean spectral density function (SDF) by frequency, of the AR(4) time series component $\{\nu_t\}$ (black lines), along with simultaneous 95% credible intervals (gray regions).

ible intervals for the temperature anomalies. We use the `simconf.mc` function from the `excursions` R packages to calculate these simultaneous intervals (see Bolin and Lindgren, 2015, 2017, 2018, for further details). Figure 2(b) displays the posterior standard deviation of the global temperature anomalies by year. The global temperature anomalies exhibits more uncertainty in earlier years and, over longer time scales, the uncertainty then tends to decrease. However, there are time periods that have more uncertainty than other periods, such as around and between the two World Wars, and in the last two years. We will learn later that the greater uncertainty at the beginning and in the last two years is partially due to the fact there is no data for the JMA series for these time periods. The uncertainty around the two wars is a feature of most data products, and is due to vastly reduced commercial shipping. The posterior mean latent global temperature anomaly shows a strong systematic and nonlinear trend, and seems to exhibit significant dependence over time.

To further explore the time series structure of the global temperature anomaly, Figure 3(a) shows posterior summaries of the trend component $\{\mu_t\}$ and the spectral density function (SDF) of the time series component $\{\eta_t\}$ in Figure 3(b). In each panel the black line denotes the posterior mean calculated for each year, and the gray regions are simultaneous 95% credible intervals again calculated using the `excursions` R package.

As explained above, we use $b = 5$ b-spline basis functions to capture long term trends on a scale of approximately 25 years in the latent climate anomaly. Over this time scale, Figure 3(a) shows that the temperature anomaly tends to decrease from 1880 to 1900, and then increases from 1900 to 2020, but the rate of increase is not the same over this period. There is a steady increase from 1900 to the mid 1950s, a slower increase until the 1980s, and then the rate of increase is highest from 1980 to 2020. Since 1980, we estimate that the global temperature anomaly has increased on average by 0.020 degrees Celsius per year.

To understand the variation of the global anomaly over shorter time scales we look at the SDF. The SDF is the Fourier transform of the autocovariance of the time series, and

allow us to decompose the variation of the time series component in terms of the variance contributions of sinusoids over different frequencies. Figure 3(b) demonstrates that there is strong time series dependence over longer time scales, as expressed by the peak at zero frequency (which is confounded with the trend), and there is also a broadband spectral peak between frequencies 0.1 and 0.2. This indicates strong time series dependence in the global temperature anomaly on a scale of 8–10 years. (If there was no time series dependence the SDF would be constant over frequency.) We can see such quasi-periodic dependence over this time scale in Figure 2(a).

4.2 Exploring the discrepancy series

For each data product series $j = 1, \ldots, J$ included in our model, we include a discrepancy term $\{\delta_{j,t}\}$ that captures the variability inherent in each data product that is not accounted for by the natural variability of each data product. Looking at posterior summaries of these terms allows us to learn about the variations over time that are not captured by the latent global temperature anomaly or the natural variability.

Figure 4 shows the posterior mean (black lines) and pointwise 95% credible interval for each discrepancy term $\{\delta_{j,t}\}$ as we vary the data product j. While there are patterns over time, we first note that most discrepancies are small relative to the scale of the global temperature series as shown in Figure 2(a) (The 95% credible intervals for the discrepancies lies between -0.15 and 0.15 degrees Celsius). Regardless, we see differences in the mean and variability of the discrepancies over time for the different data products.

The discrepancy term for the JMA series indicates that, after accounting for the natural variability, this data product varies systematically from the consensus estimate (the global temperature anomaly) in the earlier and later time periods. As explained above, the Japanese Meteorological Office changed from using GHCN to using monthly values provided by the WMO (the coverage of GHCN is larger than the WMO data). Another difference is that the ocean temperatures are from the JMO COBE-SST data set.

The NOAA discrepancy term is more variable than the other series over the entire time period, which may indicate that the standard error reported for NOAA is either too large, or that all the standard errors reported for the other series are too small.

The Berkeley and HadCRUT5 discrepancies have some similarities over time which indicate that those two data products agree to some extent (the sample correlation between the posterior means is 0.620), but still vary from the latent global temperature anomaly. Similarly, the NOAA and GISS discrepancies, while varying in their uncertainty over time are also strongly correlated over time, with a sample correlation of 0.758 between the two posterior means. However, the discrepancy terms for Berkeley, HadCRUT5 and GISS have lower variability in recent years which also indicate less deviation relative to the latent global anomaly, after accounting for the reported error in the series.

Finally, we observed that every series has some discrepancy around World War 2, which hints to variabilities in accounting for a lack of data coverage in this time period.

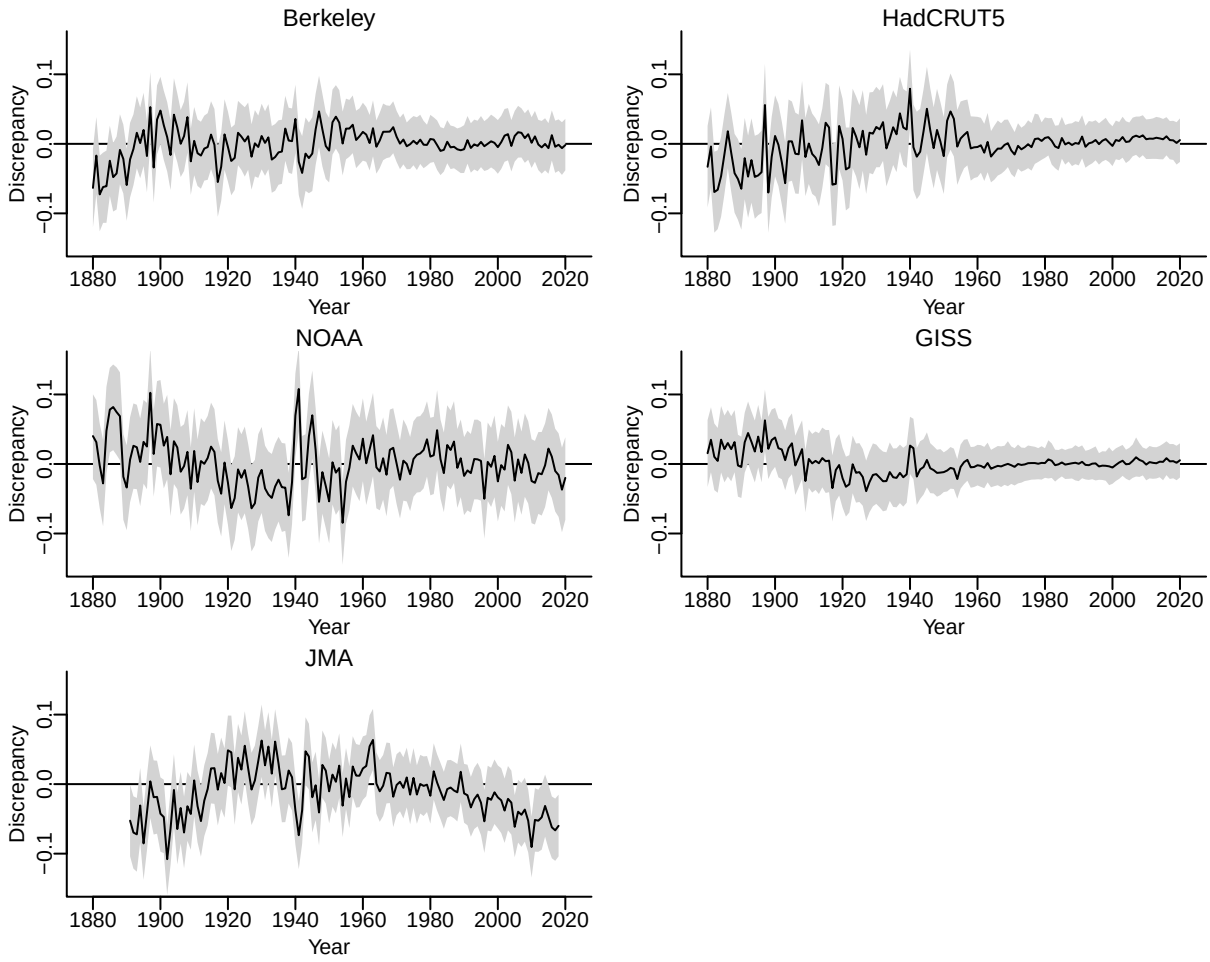

Figure 4. Posterior mean discrepancies $\{\delta_{j,t}\}$ for each data product j (black lines), along with pointwise 95% credible intervals (gray regions).

4.3 Sensitivity to observation series

Given that we are combining data products to obtain one estimate, along with a measure of uncertainty, it is also of interest to understand the influence of each data product upon the results. Complementary to the previous section, to explore this, we remove each data product from the model, and refit our hierarchical model. Figure 5 shows a plot of the posterior means (panel (a)) and posterior standard deviations (SDs) (panel (b)) by year. The different colors in each plot indicate the values when a single data product is removed as compared to the "Original" (black lines) when we do not remove a data product when fitting the model.

Figure 5 shows that some data products are influential over certain periods of year. We can see that from 1880 to around 1910, that removing GISS or JMA does change the posterior mean global anomaly over this time period. We also notice differences in the posterior mean global temperature anomaly when we remove these two data products in the period 1920–1970. Except when we remove the JMA series, in later years there is more agreement between the posterior mean global temperature anomalies when we remove a series and when we do not remove any series. This indicates good concordance in our

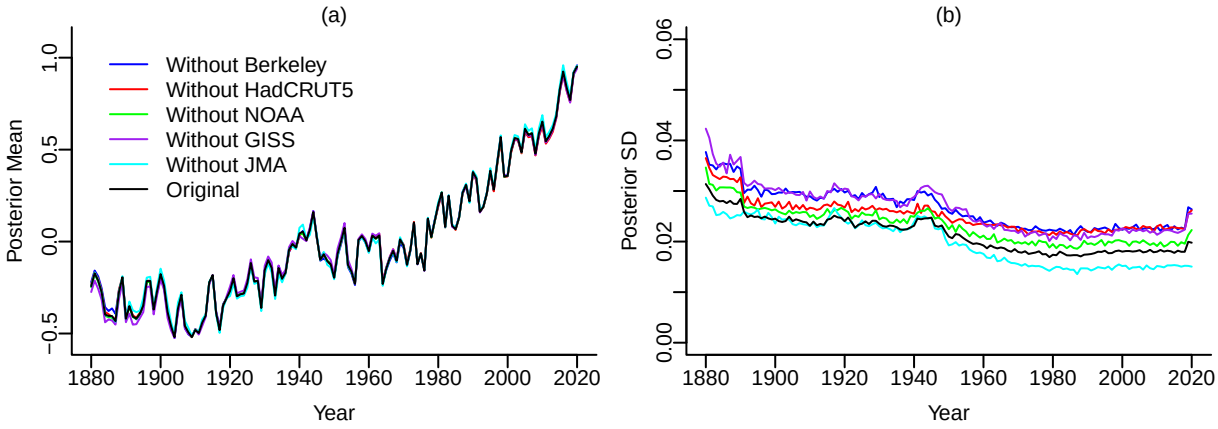

Figure 5. (a) Time series plots of the posterior mean global temperature anomalies when the hierarchical model is fit after removing each series one at a time (colored lines), as compared to the posterior mean global temperature anomalies fit using all the data (black lines). (b) Corresponding plot of the posterior SDs, by year.

estimate of the mean behavior of the latent global temperature anomaly in the last forty years. However as expected, Figure 5(b) indicates that the estimates of uncertainty vary greatly when we remove a data product from our model. Removing GISS or Berkeley greatly inflates the posterior SD for the global temperature anomaly, relative to not removing those series or removing HadCRUT5, NOAA, or JMA. Regardless, we learn that indeed the uncertainty is still larger at the start and end of the series and around the two World Wars. The flattening of the uncertainty in later years when JMA is removed indicates that JMA behaves very differently from the other series, which confirms our earlier findings.

4.4 Ranking

A feature of some interest with the public is the ranking of a given year in the temperature record, answering questions such as "Was 2020 the warmest year on record?" or "Which is the warmest decade on record?" Using the methodology of Guttorp and Kim (2013) we answer a slightly more general question, namely "What is the probability that 2020 was the warmest year on record?" Because of differences in the uncertainty assessments for the different series, we compute the probability that 2020 is the warmest year on record for most of the series used in this paper. This probability is the relative frequency of each rank among a large number of simulated temperature records from a normal distribution with the mean and variance given by the respective series and its standard error. Each calculation of the probability was based on 100 000 simulations. The study in Guttorp and Kim (2013) indicates that the dependence structure in the series hardly affects the ranks, so we simulate paths using independent normals.

Table 1 shows the estimated probabilities that 2020 is the warmest year for four global anomaly series, as well the posterior global temperature anomalies calculated via our hierarchical model using the AR(2) and AR(4) error assumption, respectively. (We cannot calculate the probability that 2020 is the warmest year for the JMA, because we do not

Table 1. For four global anomaly series and the posterior global temperature anomalies calculated via our hierarchical model using the AR(2) and AR(4) error assumption, an estimation of the probability that 2020 was the warmest year based on 100 000 simulations using the mean and variance for each anomaly series, along with the earliest year among the simulations that was the warmest year.

	$\widehat{P}$(2020 warmest)	Earliest warmest year
Berkeley	0.23	2016
HadCRUT5	0.32	2016
NOAA	0.30	2005
GISS	0.58	2016
JMA	—	2014
Posterior, AR(2) errors	0.39	2016
Posterior, AR(4) errors	0.41	2016

observe data for 2019 or 2020 for that series.) Based on 100 000 simulations, the estimated standard error for each estimated probability is less than 0.0016, smaller than the accuracy presented in Table 1. There is some variation by series in the estimated probability that 2020 is the warmest year. The probability is smallest for Berkeley (0.23), and then is similar for HadCRUT5 and NOAA (0.32 and 0.30 respectively). For GISS the probability is much larger (0.58). This is not surprising as examining Figure 1 shows that GISS has the highest average anomalies and lowest uncertainty in 2020 among all the data products. Our estimate of the probability from our model varies little: 0.39 for the AR(2) model, and 0.41 for the AR(4) model. As expected, adjusting for all the data products we estimate a probability in the middle of all the individual probabilities.

In the second column of the table we calculated the earliest year that was warmest across the simulations for each series. Except for the NOAA and JMA series, the earliest warmest year among the simulations was 2016 – this includes our posterior estimates of the global anomaly for both AR(2) and AR(4) time series model components. For the JMA the earliest warmest year was 2014 (remember there is no data for 2019 or 2020). For NOAA the earliest year among all the simulations that was warmest year was 2005: re-examining the standard errors by year (see e.g., Figure 1) we see there is a heightened uncertainty since 2004 that is not present in the other anomaly series which could account for this change.

These results indicate strong evidence of warming in recent years.

5 Discussion

Since satellite measurements of global surface temperature only have been available since 1978, and do not come with estimated uncertainties, we have not included them in our analysis. However, Figure 6 shows, in blue, the University of Alabama Huntsville satel-

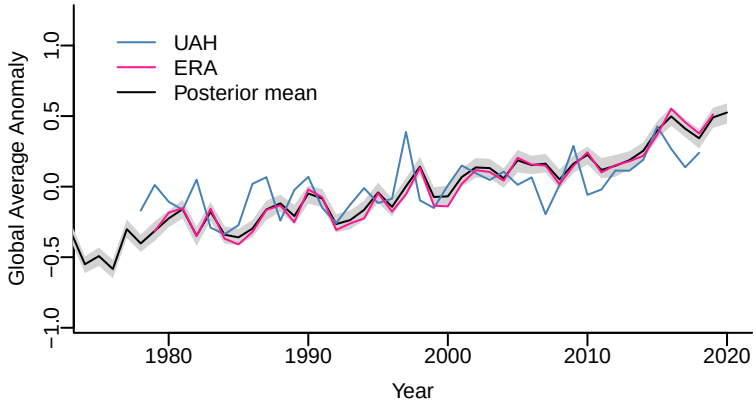

Figure 6. Posterior simultaneous 95% credibility interval for global mean annual temperature. Overlaid are the UAH satellite observations (blue) and ERA series (pink). All values are anomalized with respect to the years 1979–2018.

lite global annual means (Spencer et al., 2017) overlaid on a simultaneous 95% credibility interval again calculated using the `excursions` R package. It is important to note that, while the instrumental record is a combination mainly of 2 m temperature on land, and water temperature just under the water surface (the Berkeley team distinguishes between temperature above and below the ice surface; the difference is less than .05 °C globally), the satellite data measure the thermal microwave emission from atmospheric oxygen in the 50-60 GHz oxygen absorption complex, which is translated to a brightness temperature and then (linearly) related to the atmospheric temperature. The satellite series used corresponds to temperature in the lower troposphere, not exactly surface temperature. The satellite data tend to be colder than the instrumental data in more recent years.

It is also of interest to compare our results to the European reanalysis series ERA5, produced by the European Center for Medium-Range Forecasts (Hersbach et al., 2020). A reanalysis is a recomputation of a current weather forecast model using historical data, and is often thought of as the ground truth for climate modellers. This is shown by the pink lines in Figure 6. As can be seen, the reanalysis closely matches the posterior draws from our model, although there is a tendency in the 1980s and 1990s for the ERA to yield slightly colder global temperature anomalies in this period.

Another source of information about global temperature is the various deterministic partial differential equation models that describe the Earth system. There is a recent experiment (CMIP6, Eyring et al. (2016)) that includes runs that use historical climate forcings to estimate historical climate. In future work we will assess whether this informs our estimated global temperature.

Acknowledgments We are grateful to David Bolin, Finn Lindgren, John Kennedy, Colin Morice, Thordis Thorarinsdottir, Charlotte Wickham, and officials at the Tokyo Climate Center for helpful discussions. Data were obtained from the Copernicus Data Center, Berkeley Earth, the Goddard Institute for Space Sciences, the UK Met Office Hadley Cen-

ter, NOAA's National Centers for Environmental Information, and the Tokyo Climate Center.

References

Bolin, D. and Lindgren, F. (2015). Excursion and contour uncertainty regions for latent Gaussian models. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 77:85–106. 10

Bolin, D. and Lindgren, F. (2017). Quantifying the uncertainty of contour maps. *Journal of Computational and Graphical Statistics*, 26:513–524. 10

Bolin, D. and Lindgren, F. (2018). Calculating probabilistic excursion sets and related quantities using excursions. *Journal of Statistical Software*, 86:1–20. 10

Brockwell, P. J. and Davis, R. A. (2002). *Introduction to Time Series and Forecasting (Second Edition)*. Springer-Verlag, New York, NY. 8

Craigmile, P. F. and Guttorp, P. (2019). Modeling and assessing climatic trends. In Gelfand, A., Fuentes, M., Hoeting, J., and Smith, R., editors, *Handbook of Environmental and Ecological Statistics*. Chapman and Hall/CRC, New York: NY. 9

Das, B. (2000). *Global covariance modeling: a deformation approach to anisotropy*. PhD thesis, University of Washington. 7

Eyring, V., Bony, S., Meehl, G. A., Senior, C. A., Stevens, B., Stouffer, R. J., and Taylor, K. E. (2016). Overview of the Coupled Model Intercomparison Project phase 6 (CMIP6) experimental design and organization. *Geosci. Model Dev.*, 9:1937–1958. 15

Freeman, E., Woodruff, S., Worley, S., Lubker, S., Kent, E., Angel, W., Berry, D., Brohan, P., Eastman, R., Gates, L., Gloeden, W., Ji, Z., Lawrimore, J., Rayner, N., Rosenhagen, G., and Smith, S. (2017). ICOADS Release 3.0: A major update to the historical marine climate record. *Int. J. Climatol. (CLIMAR-IV Special Issue)*, 37:2211–2237. 4

Gneiting, T. (2013). Strictly and non-strictly positive definite functions on spheres. *Bernoulli*, 19:1327–1349. 6

Guttorp, P. and Kim, T. Y. (2013). Uncertainty in ranking the hottest years of US surface temperatures. *Journal of climate*, 26:6323–6328. 13

Hansen, J., Ruedy, R., Sato, M., and Lo, K. (2010). Global surface temperature change. *Rev. Geophys.*, 48. 5

Hegerl, G. C., Karl, T. R., Allen, M., Bindoff, N. L., Gillett, N., Karoly, D., Zhang, Z., and Zwiers, F. (2006). Climate change detection and attribution: Beyond mean temperature signals. *J. Clim.*, 19:5058–5077. 4

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horanyi, A., Munoz-ĂŘSabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., Chiara, G. D., Dahlgren, P., Dee, D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hogan, R. J., Holm, E., Janiskova, M., Keeley, S., Laloyaux, P., Lopez, P., Lupu, C., Radnoti, G., de Rosnay, P., Rozum, I., Vamborg, F., Villaume, S., and Thepaut, J. (2020). The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146:1999–2049. 15

Huang, B., Menne, M. J., Boyer, T., Freeman, E., Gleason, B. E., Lawrimore, J. H., Liu, C., Rennie, J. J., Schreck, C. J., Sun, F., Vose, R., Williams, C. N., Yin, X., and Zhang, H. (2019). Uncertainty estimates for sea surface temperature and land surface air temperature in noaaglobaltemp version 5. *J. Climate*. 5

Huang, B., Thorne, P., Banzon, V., Boyer, T., Chepurin, G., Lawrimore, J., Menne, M., Smith, T., Vose, R., and Zhang, H.-M. (2017). Extended reconstructed sea surface temperature, version 5 (ERSSTv5): Upgrades, validations, and intercomparisons. *Journal of Climate*, 30:8179–8205. 5

Ishii, M., Shoji, A., S, S., and Matsumoto, T. (2005). Objective analyses of sea-surface temperature and marine meteorological variables for the 20th century using ICOADS and the Kobe Collection. *Int. J. Climatol.*, 25:865–879. 5

Kennedy, J. J., Rayner, N. A., Atkinson, C. P., and Killick, R. E. (2019). An ensemble data set of sea surface temperature change from 1850: the met office hadley centre hadsst.4.0.0.0 data set. *Journal of Geophysical Research: Atmospheres*, 124. 5

Lenssen, N., Schmidt, G., Hansen, J., Menne, M., Persin, A., Ruedy, R., and Zyss, D. (2019). Improvements in the gistemp uncertainty model. *J. Geophys. Res. Atmos.*, 124:6307–6326. 5

Menne, M., Williams, C., Gleason, B., Rennie, J., and Lawrimore, J. (2018). The global historical climatology network monthly temperature dataset, version 4. *J. Climate*. 4

Morice, C. P., Kennedy, J. J., Rayner, N. A., Winn, J. P., Hogan, E., Killick, R. E., Dunn, R. J. H., Osborn, T. J., Jones, P. D., and Simpson, I. R. (2020). An updated assessment of near-surface temperature change from 1850: the HadCRUT5 dataset. *Journal of Geophysical Research: Atmospheres*. Available from: https://doi.org/10.1029/2019JD032361. 5

Rennie, J. J., Lawrimore, J. H., Gleason, B. E., Thorne, P. W., Morice, C. P., Menne, M. J., Williams, C. N., de Almeida, W. G., Christy, J., Flannery, M., Ishihara, M., Kamiguchi, K., Klein Tank, A. G., Mhanda, A., Lister, D. H., Razuvaev, V., Renom, M., Rusticucci, M., Tandy, J., Worley, S. J., Venema, V., Angel, W., Brunet, M., Dattore, B., Diamond, H., Lazzara, M. A., Le Blancq, F., Luterbacher, J., Mächel, H., Revadekar, J., Vose, R. S., and Yin, X. (2014). The international surface temperature initiative global land surface databank: monthly temperature data release description and methods. *Geosci. Data J.*, 1:75–102. 4

Rohde, R., Muller, R., Jacobsen, R., Perlmutter, S., Rosenfeld, A., Wurtele, J., Curry, J., Wickham, C., and Mosher, S. (2013). Berkeley Earth temperature averaging process. *Geoinfor Geostat: An Overview*. 5

Rohde, R. A. and Hausfather, Z. (2020). The berkeley earth land/ocean temperature record. *Earth Syst. Sci. Data Discuss.* Available from: https://doi.org/10.5194/essd-2019-259. 5

Spencer, R., Christy, J., and Braswell, W. (2017). Uah version 6 global satellite temperature products: Methodology and results. *Asia-Pacific Journal of Atmospheric Sciences*, 53:121–130. 15

Vose, R., Arndt, D., Banzon, V., Easterling, D., Gleason, B., Huang, B., Kearns, E., Lawrimore, J., Menne, M., Peterson, T., Reynolds, R., Smith, T., Williams Jr, C., and Wuertz, D. (2012). NOAA's merged land-ocean surface temperature analysis. *Bulletin of the American Meteorological Society*, 93:1677–1685. 5

Zhang, H.-M., Lawrimore, J. H., Huang, B., Menne, M. J., Yin, X., Sanchez-Lugo, A., Gleason, B. E., Vose, R., Arndt, D., Rennie, J. J., and Williams, C. N. (2019). Updated temperature data give a sharper view of climate trends. *Eos*, 100. Available from: https://doi.org/10.1029/2019EO128229. 5

A The full conditional distributions used in the Markov chain Monte Carlo algorithm

We now present the steps required to carry out the MCMC algorithm. First we write the model in matrix-vector form. We have that each global temperature data product series of length N, $\boldsymbol{D}_j$, $(j = 1, \ldots, J)$ satisfies

$$\boldsymbol{D}_j \;=\; \boldsymbol{Y} + \boldsymbol{\delta}_j + \boldsymbol{\epsilon}_j.$$

where $\boldsymbol{\delta}_j \sim N(\boldsymbol{0}, \tau^2 \boldsymbol{I})$ and $\boldsymbol{\epsilon}_j \sim N(\boldsymbol{0}, \boldsymbol{V}_j)$, assuming independence between these components. Then, the latent average global temperature series is

$$\boldsymbol{Y} \;=\; \boldsymbol{\mu} + \boldsymbol{\nu},$$

where the mean is $\boldsymbol{\mu} = \boldsymbol{X}\boldsymbol{\beta}$ and $\boldsymbol{\nu}$ is a realization of an AR(p) process with PACF coefficient p-vector $\boldsymbol{\psi}$ and innovation variance σ^2. (Here $\boldsymbol{X}$ is an $N \times b$ matrix with tth row given by $\boldsymbol{x}(t)$.) Let $\boldsymbol{\Sigma} = \mathrm{cov}(\boldsymbol{\nu})$ be the covariance matrix for $\boldsymbol{\nu}$, which we parameterize in terms of $\boldsymbol{\eta}$ and σ^2. Then our hierarchical model can be expressed as

$$\begin{aligned} \boldsymbol{D}_j | \boldsymbol{Y}, \tau^2 &\sim N(\boldsymbol{Y}, \tau^2 \boldsymbol{I} + \boldsymbol{V}_j), \quad j = 1, \ldots, J; \\ \boldsymbol{Y} | \boldsymbol{\beta}, \boldsymbol{\eta}, \sigma^2 &\sim N(\boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{\Sigma}). \end{aligned}$$

We now present the full conditional distributions and the MCMC updates for each parameter in the model.

Updating $\boldsymbol{Y}$: We have

$$\pi(\boldsymbol{Y} | \{\boldsymbol{D}_j\}, \boldsymbol{\theta} \setminus \{\boldsymbol{Y}\}) \;\propto\; \left[\prod_{j=1}^{J} f(\boldsymbol{D}_j | \boldsymbol{Y}, \tau^2) \right] \pi(\boldsymbol{Y} | \boldsymbol{\mu}, \boldsymbol{\Sigma}).$$

Thus, we sample $\boldsymbol{Y}$ from $N(\boldsymbol{P}^{-1}\boldsymbol{q}, \boldsymbol{P}^{-1})$ where

$$\begin{aligned} \boldsymbol{P} &= \sum_{j=1}^{J} \boldsymbol{B}_j^* + \boldsymbol{\Sigma}^{-1}; \\ \boldsymbol{q} &= \sum_{j=1}^{J} \boldsymbol{B}_j^* \boldsymbol{D}_j^* + \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}, \end{aligned} \tag{A.1}$$

where $\boldsymbol{B}_j^*$ is an $N \times N$ matrix with (t, t') element

$$B_{j,t,t'}^* \;=\; \begin{cases} B_{j,t,t'}, & o_{j,t} = 1 \text{ and } o_{j,t'} = 1; \\ 0, & \text{otherwise,} \end{cases}$$

with $\boldsymbol{B}_j = \left[\tau^2 \boldsymbol{I} + \boldsymbol{V}_j\right]^{-1}$, and $\boldsymbol{D}_j^*$ is an N-vector with elements $(t = 1, \ldots, N)$

$$D_{j,t}^* \;=\; \begin{cases} D_{j,t}, & o_{j,t} = 1; \\ 0, & \text{otherwise.} \end{cases}$$

Updating τ^2 : We have

$$\pi(\boldsymbol{Y}|\{\boldsymbol{D}_j\}, \boldsymbol{\theta} \setminus \{\tau^2\}) \quad \propto \quad \left[\prod_{j=1}^{J} f(\boldsymbol{D}_j|\boldsymbol{Y}, \tau^2)\right] \pi(\tau^2).$$

A Metropolis-Hastings symmetric random walk update is used on the parameter $\log \tau^2$. Suppose we are at $\log \tau^2$ and we propose an update from $\log \tau^2_{new} \sim N(\log \tau^2, \kappa^2)$ for some constant positive variance κ^2. Including the Jacobian of the log transformation for τ^2, the new value is accepted with probability $\min(\exp(\alpha), 1)$ where

$$\begin{aligned} \alpha \quad = \quad & \sum_{j=1}^{J} \log f(\boldsymbol{D}_j|\boldsymbol{Y}, \tau^2_{new}) + \log \pi(\tau^2_{new}) + \log \tau^2_{new} - \\ & \sum_{j=1}^{J} \log f(\boldsymbol{D}_j|\boldsymbol{Y}, \tau^2) - \log \pi(\tau^2) - \log \tau^2. \end{aligned}$$

In calculating α we only calculate $f(\boldsymbol{D}_j|\boldsymbol{Y}, \tau^2_{new})$ and $f(\boldsymbol{D}_j|\boldsymbol{Y}, \tau^2)$ for each $j = 1, \ldots, J$ at the subset of time points for which data is available (i.e., for those values of t such that $o_{j,t} = 1$).

Updating $\boldsymbol{\beta}$: We have

$$\pi(\boldsymbol{Y}|\{\boldsymbol{D}_j\}, \boldsymbol{\theta} \setminus \{\boldsymbol{\beta}\}) \quad \propto \quad \pi(\boldsymbol{Y}|\boldsymbol{\mu}, \boldsymbol{\Sigma})\pi(\boldsymbol{\beta}),$$

and hence we sample $\boldsymbol{\beta}$ from $N(\boldsymbol{P}^{-1}\boldsymbol{q}, \boldsymbol{P}^{-1})$ where

$$\boldsymbol{P} \; = \; \boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X} + \boldsymbol{V}_{\beta}^{-1}; \quad \text{and} \quad \boldsymbol{q} \; = \; \boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{Y} + \boldsymbol{V}_{\beta}^{-1}\boldsymbol{m}_{\beta}.$$

Updating $\boldsymbol{\eta}$ and σ^2: We have

$$\pi(\boldsymbol{Y}|\{\boldsymbol{D}_j\}, \boldsymbol{\theta} \setminus \{\boldsymbol{\eta}, \sigma^2\}) \quad \propto \quad \pi(\boldsymbol{Y}|\boldsymbol{\mu}, \boldsymbol{\Sigma})\pi(\boldsymbol{\eta})\pi(\sigma^2).$$

A Metropolis-Hastings symmetric random walk update is used on the parameters $\boldsymbol{\eta}$ and $\log \sigma^2$. Suppose that we are currently at $\boldsymbol{\eta}$ and $\log \sigma^2$. We then propose

$$\boldsymbol{\eta}^{new} \sim N_p(\boldsymbol{\eta}, \boldsymbol{K}) \quad \text{and} \quad \log \sigma^2_{new} = N(\log \sigma^2, \kappa^2),$$

for some positive definite matrix $\boldsymbol{K}$ and positive variance κ^2. These new values are accepted with probability $\min(\exp(\alpha), 1)$ where

$$\begin{aligned} \alpha \quad = \quad & \log \pi(\boldsymbol{Y}|\boldsymbol{\mu}, \boldsymbol{\Sigma}^{new}) + \log \pi(\boldsymbol{\eta}^{new}) + \log \pi(\sigma^2_{new}) + \log \sigma^2_{new} - \\ & \log \pi(\boldsymbol{Y}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) - \log \pi(\boldsymbol{\eta}) - \log \pi(\sigma^2) g1 - \log \sigma^2, \end{aligned}$$

where $\boldsymbol{\Sigma}^{new}$ is the covariance matrix for $\boldsymbol{Y}$ calculated with $\boldsymbol{\eta}^{new}$ and $\log \sigma^2_{new}$ and $\boldsymbol{\Sigma}$ is the covariance matrix for $\boldsymbol{Y}$ calculated with $\boldsymbol{\eta}$ and $\log \sigma^2$.

Sampling the data discrepancies: In our model, $\boldsymbol{\delta}_j$ is the discrepancy for data product series $j = 1, \ldots, J$. We evaluate the posterior distribution of discrepancies using our posterior draws of the parameters $\boldsymbol{\theta}$: for each j we have

$$\boldsymbol{\delta}_j | \boldsymbol{D}_j, \boldsymbol{\theta} \sim N(\boldsymbol{P}^{-1}\boldsymbol{q}, \boldsymbol{P}^{-1}),$$

where $\boldsymbol{P} = \boldsymbol{I}/\tau^2 + \boldsymbol{V}_j^{-1}$ and $\boldsymbol{q} = (\boldsymbol{D}_j - \boldsymbol{Y})/\tau^2$.