Atmospheric Measurement Techniques Discussions

Atmos. Meas. Tech. Discuss., referee comment RC1
https://doi.org/10.5194/amt-2020-454-RC1, 2021

Comment on amt-2020-454

Anonymous Referee #2

Referee comment on "Global ensemble of temperatures over 1850–2018: quantification of uncertainties in observations, coverage, and spatial modeling (GETQUOCS)" by Maryam Ilyas et al., Atmos. Meas. Tech. Discuss., https://doi.org/10.5194/amt-2020-454-RC1, 2021

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

Maryam Ilyas, Douglas Nychka, Chris Brierley, and Serge Guillas

An interesting paper with a few issues to be resolved. In their analysis of global temperature data using the multiresolution lattice kriging method, the authors extend the work of Ilyas et al 2017 by exploring the hyperparameter estimation uncertainty using a Monte Carlo sampling method. It is interesting to see the impacts of this hyperparameter uncertainty assessment. These are a potentially important source of uncertainty in assessments of observed global temperature change that have not previously been investigated in other studies. There are some issues with the paper structure, including a lack of concluding remarks. Additional discussion of the effects controlled by the sampled parameters and illustration of the impacts of their sampling throughout the temperature record is needed.

Main points:

1. While it is great that the paper includes estimates of hyperparameter uncertainty, I'm am left uncertain on the extent to which hyperparameter uncertainty translates into an appreciable uncertainty in the temperature fields and how this varies through the temperature record. The paper only provides examples of hyperparameter estimates and resulting fields for a single month. Is this representative of other months? Interpretation of differences between the ABC based and profile likelihood based analyses would be aided by showing how the profile likelihood hyperparameter estimates compare to those from

ABC, for example in Figure 2. Similarly, temperature/uncertainty fields are only shown for a single month in Figures 3/4.

These points seem important to understand the benefits of sampling the hyperparameter uncertainty, also given that the method appears to be computationally expensive.

2. The review of prior literature is frequently a few years out of date and needs updating. Some cited studies are inaccurately or incorrectly described. See detailed comments for details.

3. There appear to be a few simplifications in the statistical model/uncertainty model used that are not discussed:

- Hyperparameter estimates are global, estimated independently for each field, with no regional estimates, essentially modelling temperature anomaly variability as being identical at all locations over land and sea.
- As a space only model (not space-time) there appears to be no accounting for persistence of temperatures used to aid reconstruction or accounted for in uncertainty estimates.
- My understanding of MRLK is that it models observational error as identically distributed for each observed location. The analysis described makes no mention of the additional uncertainty terms for HadCRUT4 (in addition to the ensemble) that describe differences in observational error distributions for each grid cell and correlations in errors between grid cells, arising from the movement of marine measurement platforms. It appears that this information, that is not encoded into the HadCRUT4 ensemble members, has not been used and they are not described in Section 3.2. These uncertainties were found to be important in Morice et al 2012. Some comment on not including these, or how they are approximated by the additive uncorrelated error term in MRLK, would be appropriate.

4. The reader needs to refer to Nychka et al 2015 to understand the meaning of the lambda hyperparameter. The "aw" hyperparameter (please rename to use a single letter unless it is a product of two variables) does not appear to be described in Nychka et al 2015. I do not understand how to interpret the function of this "autoregressive weights" hyperparameter and how it might affect the resulting temperature fields.

5. Discussion of Figure 5 suggests that uncertainty estimates for global average temperature anomalies are wider for ABC that those of Ilyas et al 2015, but this is not particularly evident in Figure 5. Uncertainty ranges appear to be roughly comparable, of similar magnitude to the quantisation in the plot of roughly 0.01 °C, but slightly skewed one way or another. It's not clear that the ABC sampling of the hyperparameter uncertainty would necessarily lead to wider uncertainty estimates in the global mean than the profile likelihood estimates. For example, the lambda parameter is defined in Nychka

et al 2015 as lambda =noise variance / process variance. Sampling into high values of lambda would give a process with low variance and large measurement noise, which would lead to smaller uncertainties arising from sampling limitations. The lower variance of the ABC analysis field in Figure 3 suggests that this might be the case. It would be an alternative explanation to the changes in LatticeKrig 6.4 that are alluded to in the first paragraph on page 10.

6. The paper ends rather abruptly with a discussion of a sampling method (which arguably should be moved earlier in the paper). It would benefit from a conclusions section. Are there any deficiencies in the approach that we should be aware of? What's missing or could further developed? It could comment on developments while this paper was being worked on that are not included, e.g. for HadSST4 and talk more broadly about where this study fits alongside other research in the subject area. It would be an appropriate place to place a link to the data.

Detailed points.

Page 1, Abstract, line 1: Needs the word global in there to indicate that we're talking about global temperature records?

Page 1, Abstract, Line 7: It's not clear in the abstract what the "variation" in parameters is referring to. Hyperparameter estimates vary from month to month but not spatially. Or is this referring to the uncertainty in hyperparameter values, which otherwise isn't stated in the abstract and is the key addition in the paper?

Page 1, line 15 – Good et al 2016 is a satellite-based skin temperature record, not air temperature?

Page 1, line 19 – It's not exactly so simple as obtaining data from the WMO/GCOS. Modern messages are transmitted via these means but much work is required to compile observations from individual nation states and from research institutions to compile the historical records.

Page 2, line 11 – The most recent version of the NOAA record is now called NOAAGlobalTemp with the following reference:

Zhang, Huai-Min, Jay H. Lawrimore, Boyin Huang, Matthew J. Menne, Xungang Yin, Ahira Sánchez-Lugo, Byron E. Gleason, Russell Vose, Derek Arndt, J. Jared Rennie, Claude N. Williams, 2019: Updated temperature data give a sharper view of climate trends, Eos, 100, doi.org/10.1029/2019EO128229. Published on 19 July 2019.

A newer version is also just been published:

Vose, R. S., Huang, B., Yin, X., Arndt, D., Easterling, D. R., Lawrimore, J. H., et al. (2021). Implementing full spatial coverage in noaa's global temperature analysis. Geophysical Research Letters, 48, e2020GL090873. https://doi.org/10.1029/2020GL090873

Page 2, line 14 – Ishii et al., 2005 describes only the marine portion of the JMA temperature record.

Page 2, line 15 – The latest version of HadCRUT, HadCRUT5, has the following reference. Note the date for the final published version as 2021, not 2020:

Morice, C. P., Kennedy, J. J., Rayner, N. A., Winn, J. P., Hogan, E., Killick, R. E., et al. (2021). An updated assessment of nearâ□□surface temperature change from 1850: the HadCRUT5 data set. Journal of Geophysical Research: Atmospheres, 126, e2019JD032361. https://doi.org/10.1029/2019JD032361

Page 2, line 15 – The 2013 paper for Berkeley Earth only described the land data. The recent paper describing the merged land-ocean can be cited as:

Rohde, R. A. and Hausfather, Z.: The Berkeley Earth Land/Ocean Temperature Record, Earth Syst. Sci. Data, 12, 3469–3479, https://doi.org/10.5194/essd-12-3469-2020, 2020.

Page 2, line 19 – The GISS data set has long only used satellite nightlight data for bias adjustment of urban areas. It does not use satellite derived temperature information. The current version of MLOST does not use satellite data. The statements here likely refer to the ERSST3b sea surface temperature data set, which used satellite data and was once the marine data source for these data sets. The current version, ERSST5, does not use satellite data.

Page 2, line 26 – HadCRUT4 is not interpolated but the recently publish HadCRUT5 is. The JMA data set's oceans are interpolated.

Page 2, line 30 – MLOST is not based on linear interpolation. It's a combination of "low frequency" spatial running average and a "high frequency" reduced space analysis using a

method called Empirical Orthogonal Teleconnections.

Page 2 line 31 – GISS uses linear distance weighting, not inverse linear distance weighting. No inverse involved (see Equation 1 of Lenssen et al., 2019 for the equation, or section 2 of Hansen et al., 2010 for a description). The linear distance weighting is correctly stated in the following sentence on line 32.

Page 2, line 34 – A reference is needed here. Does this use of kriging refer to the JMA COBE SST data set's use of optimum interpolation?

Page 3, line 1 – Cowtan and Robert (2014) should be Cowtan and Way (2014) (i.e. the author's name is Robert Way). Repeated again in other references to this paper (e.g. on line 3).

Page 3, line 4 – I would argue that these methods do not ignore variations at multiple length scales. For kriging/Gaussian process regression, the ability to represent multiple length scales is dependent on the covariance function used (which can be extremely flexible if constructed to be). The reconstruction method in NOAAGlobalTemp also represents multiple length scales in its own way though a reduced space decomposition. I assume that this comment on multiple length scales is alluding to MLRK, which explicitly represents multiple length scales as a sum of covariance functions. The distinction here is perhaps in MRLK flexibly to fit covariance structures with multiple scales without necessarily defining those structures in advance?

Page 3, line 5 – The new HadCRUT5 data set include a conditional simulation step to encode analysis uncertainties into an ensemble. Other data sets provided uncertainty estimates in their interpolation by other means but not through simulation.

Paragraph at page 3, line 15 - It is great to see the hyperparameter uncertainties and conditional simulation included. It does not appear that all components of the HadCRUT4 uncertainty model have been used though. In particular, those associated with biased observations for individual ships encoded into the HadCRUT4 error covariance matrices and per-grid-cell uncertainty estimates, not included in the ensemble, are not used. Instead the model seems to assume i.i.d. errors for each observed grid cell, with a stationary measurement error variance across all locations, estimated each month.

Page 3, line 27 – Should this comment on sparsity in covariance matrices refer instead to sparsity in inverse covariance matrices?

Page 4, line 19 – I know that I can refer to Nycha et al. (2015) to understand the function

of the lambda parameter. The average reader will not know this. A reference to Nycha et al. (2015) would be appropriate here. Is "aw" one term or two? I don't understand what this parameter does and I can't find it in Nycha et al.

Page 4, line 21 – Typo? "The smoothness parameter lambda influence throughout the calculation". What does is it influence?

Page 5, line 12 – Are these semivariances defined at the observation locations. Is there any binning etc. to compensate for biases sampling of e.g. short ranges when computing the empirical semivariogram?

Page 5, line 4 – d and rho are not defined in this paper.

Page 5, line 10 – version 4.5.0.0 but it's appropriate for comparisons with Ilyas et al. (2017) if it is the same version as used there.

Page 7, line 6 – This Section 3.2 is essentially a recap of Morice et al (2012), with a heavy focus on the land data. Only the large-scale bias terms are discussed here and not the measurement and grid sampling uncertainty components. These are particularly important for marine regions as ship/buoy movement leads to spatially correlated error, which should be mentioned here. HadCRUT4 does not include these in the ensemble, but instead as additional spatial error covariance matrices. It seems that these have not been used in this paper.

Page 7, line 3 – The sentence should not begin with "So".

Page 7, line 24 – It could be noted that the error model here represents the effects of potential residual biases when using station records that have been screened for urbanisation.

Page 7, line 28 - Sampling distributions for the HadCRUT4 ensemble are described in Morice et al. (2012). It would be sufficient here to refer to that study for the ensemble sampling methods rather than repeating it here and elsewhere in Section 3.2. The output of Morice et al. (2012) is used in this paper rather than reimplementation/modification of their methods so these methodological details are not core to this study.

Page 8, line 4 – Again, sampling distributions used in the construction of the input datasets used could be replaced with a reference to Morice et al. (2012) as they are not

critical to the new work undertaken in this study.

Page 8, line 20 – Estimation of hyperparameters for each individual choice is an important design choice. There is no discussion of variation of the parameter estimates seasonally or in time later in the paper. It would be interesting to see this.

Page 8, line 24 – this is 10 hyperparameter sample draws for each month, yes?

Page 28, line 28 – do these marginal variances have interpretable units? Are these °C^2? It's interesting if the marginal variance has little affect on uncertainty as it controls the variance of the process in interpolated regions and the relative importance of each spatial scale. Or is the uncertainty in process variances somehow pushed into lambda in the model's parameterisation?

The next sentence says that this parameter is computed from a single field. Is there no seasonal variability in marginal spatial variance? Perhaps some comment on how the parameter should be interpreted would be helpful to explain why February 1988 is representative of the whole data set.

Page 9, line 2 – Use of the field with minimum coverage seems a strange choice rather than using a well sampled period. Is this because of computational cost limitations?

Figure 2 - It would be interesting to see the likelihood-based estimate here too. This would help to understand what's happening in Figure 3 in comparisons between ABC and likelihood-based fields. This figure would benefit from some accompanying discussion of the effects of the parameters on the fitted fields and how the parameter estimates differ from the likelihood-based parameters.

For example, Nychka et al (2015) indicates that lambda =noise variance / process variance. High values of lambda would give a process with low variance and large measurement noise. Would this result in e.g. lower variance fields than a likelihood based estimate of a lower lambda?

The aw parameter is interesting here. There's a lot of weight right at the edge of the prior distribution. Is the distribution being truncated by the choice of prior?

Page 9, line 11 – Some discussion of how parameters compare between ABC and likelihood methods would again be useful here. What is it about the sampled parameter

values that leads to the differences?

Page 10, line 2 – The lower uncertainty in unobserved grid locations could explain why This could be the reason that global average statistics do not appear to be much affected. It seems like the ABC ensemble is leaning towards a model with greater observational noise variance and lesser process variance. This would explain the smoother fields in the figure 2 (a)-(b) comparison. Is this correct?

Figure 3 – Are these fields the ensemble means/medians? This is not stated. Are the results for this month representative of other months in terms of parameter estimates and uncertainty estimates? It would be useful to see how they compare for better sampled periods or other times of year. The plot needs units (°C?).

Figure 4 – How is uncertainty defined here? Is this the full ensemble spread? What is the statistic being shown? What are the units?

Page 12, line 5 – I think that this says that samples are drawn from the conditional normal distribution, with each HadCRUT4 ensemble member having 10 hyperparameter samples, and each of those having 100 samples from the conditional normal. However, the wording of "namely the variogram based ABC posteriors of autoregressive weights and smoothing parameter" does not include the conditional normal sampling.

Figure 5 – Is there a grey line plotted for the ensemble median of the old ensemble? If so then I can't see it. It would be helpful to include to show any differences (or lack thereof) in the mean/median with the hyperparameter sampling.

Figure needs units (°C?).

Page 14, line 5 – It's not clear that the uncertainties are always larger for ABC. They seem comparable or slightly skewed relative to Ilyas et al 2017. Differences often appear to be around the scale of the apparent rounding resolution in the plot. It looks like the uncertainty range is often narrower for the new ensemble.

Is it guaranteed that the uncertainty estimates would be wider using ABC? Could lower uncertainties be possible if the hyperparameter distribution samples a region of the parameter space that leads to smaller process variance, and hence smaller coverage uncertainty estimates?

Page 14, paragraphs at lines 13 and 24 – This discussion of Latin hypercube sampling is a strange and abrupt way to end the paper. It would better be placed earlier in the method/results. This sampling is good to see though as a 100,000-member gridded dataset would be rather unwieldy to use.

Page 15, line 1 - No conclusions section? See main point 6.

Figure 6 – The axes need labels/units. It could be useful to see the resulting sampling for other locations too.