Geoscientific Model Development Discussions

Geosci. Model Dev. Discuss., referee comment RC2
https://doi.org/10.5194/gmd-2021-39-RC2, 2021

Comment on gmd-2021-39

Kai-Lan Chang (Referee)

Referee comment on "Calibrating a global atmospheric chemistry transport model using Gaussian process emulation and ground-level concentrations of ozone and carbon monoxide" by Edmund Ryan and Oliver Wild, Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2021-39-RC2, 2021

Even if I have a general appreciation about the content of the paper, however I think more work is needed in order to convince the reader about the result or utility of the proposal with respect to the existing literature, but I think they are addressable, though it will take some work.

Major comments:

- When I read the abstract and introduction, I have an impression that the author will apply the calibration framework developed by Kenndy and O'Hagan (2001), but it turns out that it is not the case. The author should have explained in the beginning why they do not include model discrepancy in the methodology or GP model. It is well known that model discrepancy is very important for model calibration (Brynjarsdóttir and O'Hagan, 2014), and it is obvious there is no CTM that can perfectly predict the true process, even with the optimised inputs.
  Brynjarsdóttir, J. and O'Hagan, A.: Learning about physical parameters: The importance of model discrepancy, Inverse Problem, 30, 114007, 2014

- I used to think the calibration won't work on tropospheric ozone, since accurate tropospheric ozone simulation relies on accurate regional emission inventory and meteorology. These inputs need to be continuously updated, rather than calibrated. As far as I am aware, the existing calibration approaches assume the parameters are "fixed initial condition", rather than time-varying or spatial-varying settings. Would it be making more sense to consider such input, e.g. emission inventory, as a time series?

Morris, M. D. (2012). Gaussian surrogates for computer models with time-varying inputs and outputs. Technometrics, 54(1), 42-50.

- The authors split the total output into 6528 separated GP emulators (p13). Since each emulator is treated separately, how can they be certain all the emulators yield the same calibrated inputs?
  - what if the calibrated inputs from different emulators are diverse?
  - even if the calibrated inputs from different emulators are similar, how the optimised inputs are determined?
- Given the fact that the authors build an emulator separately for each location and month, I am actually concerned whether this study is actually useful. If the inputs are only calibrated at a specific month/location, and might not be valid for other months/locations, then this message is not really useful for modelers. If they can not run the model at a single setting of the input parameters, the calibration does not consider to be working. It is also contradictory to its title ”Calibrating a global atmospheric chemistry transport model”, because the authors have not taken any spatial and temporal correlations into account in their GP model.

- A collection of 272 locations is not strictly prohibited for GP emulation/calibration, why building 24 emulators is not an option?

- Calibration is not just about estimating the optimised inputs, but also about estimating (under reasonable assumptions) the potential simulator misspecification, or discrepancy. However, this latter estimation is not explored in the studies. This should be explored more carefully, even if the model discrepancy is not accounted for. For example, does the final calibrated model outperform than the ensemble mean output from their Latin hypercube design with respect to synthetic or real observations?

- How do the authors choose the settings of their prior distributions for the hyperparameters in GP mean and covariance functions? This is not discussed anywhere.

- The authors often attribute the variation between irregularly distributed measurements and gridded output to representative error throughout the paper. This is unfair and inadequate, because model simulated output at gridded points does not imply model output is more representative at a broader scale, but simply because the model is too limited or too coarse to represent all the fine structures.

Minor comments:

- P2, l5 and P6, l4, I understand that this is a standard statement to say poor spatial

coverage of atmospheric composition measurements, but in your case the surface ozone measurements are very dense in Europe and most of the US, compared to the 2.8 degree resolution of their model output.

- p3, l4, this reference should be updated to the most recent GBD 2019 study. 4.2m premature deaths are referring to the pm2.5 estimates. This study is about ozone, so the latest estimate of 365 thousand premature deaths is more appropriate.
- p3, l14-16, this statement seems unfair, since several attempts were already made for calibrating/emulating this type of model output. See following references:
  Chang, K. L., & Guillas, S. (2019). Computer model calibration with large nonâ□□stationary spatial outputs: application to the calibration of a climate model. Journal of the Royal Statistical Society: Series C (Applied Statistics), 68(1), 51-78.
  Couvreux, F., Hourdin, F., Williamson, D., Roehrig, R., Volodina, V., Villefranque, N. et al. (2021). Processâ□□based climate model development harnessing machine learning: I. a calibration tool for parameterization improvement. Journal of Advances in Modeling Earth Systems, 13(3), e2020MS002217.
  Guan, Y., Sampson, C., Tucker, J. D., Chang, W., Mondal, A., Haran, M., & Sulsky, D. (2019). Computer model calibration based on image warping metrics: an application for sea ice deformation. Journal of Agricultural, Biological and Environmental Statistics, 24(3), 444-463.
  Karagiannis, G., Konomi, B. A., & Lin, G. (2019). On the Bayesian calibration of expensive computer models with input dependent parameters. Spatial Statistics, 34, 100258.
  Salter, J. M., Williamson, D. B., Scinocca, J., & Kharin, V. (2019). Uncertainty quantification for computer models with spatial output using calibration-optimal bases. Journal of the American Statistical Association.
- P5, l5 and p22, l21 "require thousands of model runs" appear to be exaggerating, since the authors are aware that will be dependent on how many parameters to be turned (and they only use 80 runs).
- P5, l10, "Since the first application of emulation methods for model calibration (Kennedy and O'Hagan, 2001), [...] In this study, we apply these approaches to models of tropospheric ozone for the first time to demonstrate the feasibility of parameter estimation." This is inappropriate because the authors do not consider model discrepancy and measurement uncertainty in their emulator component (see my major comment 1).
- P5, l12, Higdon et al. (2008) should be cited, since this is the first paper successfully extending the calibration framework into the "highly multivariate output".
  Higdon, D., Gattiker, J., Williams, B., & Rightley, M. (2008). Computer model calibration using high-dimensional output. Journal of the American Statistical Association, 103(482), 570-583.
- P5, l18, "Firstly, ground-level composition measurements are usually made at a single location which may not be representative of a wider region at the grid-scale of the model. Global chemistry transport models typically have a spatial scale of the order of 100 km." This statement is somehow misleading, because dense ground based measurements (especially in Europe) reflect the local fine variations that can not be solved by coarse model resolution.
- P8, l6, As I mentioned earlier, emissions are also dynamic in time and space. So the authors should comment if these parameters only represent the initial conditions.4
- P8, l19, I believe the reference is Chang et al. (2017).
- P11, l23, There are a few alternative approaches, such as principal components (Higdon et al. 2008; Holden et al., 2015) or low rank approximations (Bayerri et al., 2007; Bowman and Woods 2016; Chang and Guillas, 2019), that are proposed to tackle high dimensional output.
  Bayarri, M., Berger, J., Cafeo, J., Garcia-Donato, G., Liu, F., Palomo, J., Parthasarathy, R., Paulo, R., Sacks, J., and Walsh, D. (2007). Computer model validation with functional output. Ann. Statist., 35(5):1874–1906
  Bowman, V. E. and Woods, D. C. (2016). Emulation of multivariate simulators using

thin-plate splines with application to atmospheric dispersion. J. Uncertnty Quant., 4(1):1323–1344.
Holden, P. B., Edwards, N. R., Garthwaite, P. H., and Wilkinson, R. D. (2015). Emulation and interpretation of high-dimensional climate model outputs. J. Appl. Statist., 42(9):2038–2055.

- P12, l16, If "where B is a p × p matrix with zeros in the off diagonals" how the different input parameters can be correlated?
- P14, l1, The emulators used in this study are not taken into account measurement uncertainty or model discrepancy, so it merely represents the "output interpolator", I do not see why the authors should report R2.