Biogeosciences Discuss.,
https://doi.org/10.5194/bg-2018-96-AC3, 2018


Biogeosciences
Discussions

# *Interactive comment on* "Linking big models to big data: efficient ecosystem model calibration through Bayesian model emulation" *by* Istem Fer et al.

**Istem Fer et al.**

fer.istem@gmail.com

Received and published: 17 May 2018

**RC 2 - The comparison between the results of the emulated and the real SIPNET show that the distributions and central moments of the posteriors are different. This is seen in:**

**a. Figure 3, where there is not "superior" approach across parameters: sometimes is R3, sometimes AAO, sometimes both R2 and R3 are equally good.**

AC - After fixing errors in our algorithm, we have repeated this test with three changes: 1) we now sample the MCMC in the iterative rounds (instead of drawing from marginal



distributions), 2) we now directly plot the MCMC samples (instead of fitted parametric distributions), 3) we now use the contaminated synthetic data. Please see our other response regarding the contamination of the synthetic dataset and Fig 1 below. In the revised figure emulator performance improved and R3 was almost always the best posterior distribution in terms of resolving the true parameter.

**b. Figure 5, where 50% of the emulated SIPNET parameters are (statistically?) different from the central moment of the distribution of the "bruteforce" model calibrations and all of the emulated estimates have substantially higher ranges.**

AC - We redrew this figure after sampling joint posterior distribution and directly plotting MCMC samples themselves instead of using the approximated parametric distributions. Please see our next response and the new figure below (Fig 2).

**Both these results suggest that some further developments have to be investigated in order to rely on posteriors from emulators. It would be key to investigate why the emulators are overall inflating uncertainty and missing the optimum in particular parameters (equifinality? Non-linearities in model functions controlled by those parameters?).**

AC - We thank the reviewer for this remark. Both reviewers' comments indeed helped us investigate further developments in our workflow and visualization. With this latest improvement/fix, the differences between emulator and bruteforce posteriors are diminished further and the emulator medians are notably closer to the optima (Fig 2 below). Our two answers to why emulators are overall inflating uncertainty are the following:

1) There is room for further improvement in the workflow. As re-sampling from the joint posterior distributions rather than the marginal distributions helped with gaining more constraint, other improvements could be thought of: e.g. adaptive sampling design could be further improved, emulator could be passed to more effective algorithms than MH-MCMC, different settings in the mlegp package could be tested to optimize the Gaussian Process (GP) fitting, a different GP package could be used/written (we added

a citation comparing Gaussian process modeling software; Erickson, Ankenman and Sanchez, 2018) etc. We discuss these in the text, and we will be actively working towards such improvements in the future.

> (before the last paragraph of section 4.6): To fit the Gaussian process models in this study, we used the $mlegp$ R-package which was found to be performing well with its default settings (Erickson, Ankenman and Sanchez, 2018). The comparison by Erickson et al. (2018) shows that there are faster (such as $laGP$) and computationally more stable (such as $GPfit$) R-packages available. However, $laGP$ performs worse than $mlegp$ unless thousands of design points are provided, and $GPfit$ is substantially slower than $mlegp$ as it is solely written in R whereas $mlegp$ is pre-compiled in C. Finally, other packages from other platforms (such as the GPy and scikit-learn modules of Python) could outperform $mlegp$ (Erickson, Ankenman and Sanchez, 2018), however, as PEcAn is mainly written in R, $mlegp$ was an adequate choice for our workflow. Overall, we note that approximation error vs clock-time trade-off is not independent of the software/code used to fit the Gaussian process model.

2) The changes to the MCMC algorithm to accommodate emulator interpolation uncertainty, which is the source of the emulators inflating the uncertainty with respect to bruteforce, is an important feature of our algorithm not something that needs to be fixed. As we do not run the full model everywhere in the parameter space, it is important that the emulator not only interpolates between the points in the response surface, but also reflects uncertainty where the model was not run (a reason why we chose gaussian process as the emulator in the first place). If more design points are added, the uncertainty reduces further as shown in Figure 7 in the manuscript. Failure to formally incorporate this interpolation uncertainty (a mistake we ourselves made early in the development of this algorithm) leads to falsely overconfident posteriors that often



exclude the 'true' value.

Equifinality is a problem for bruteforce methods as well. While it is possible that this is slightly exacerbated in the emulator approach, as there is more "wiggle room" for parameter combinations, the formal uncertainty propagation due to GP approximation errors is the main reason for higher uncertainty in emulator posteriors here.

In terms of non-linearities in model functions, GPs are well-suited for the task of emulating non-linear surfaces, and are shown to be performing well regardless of the degree of non-linearities in the fitted surface (Rasmussen, 1996; Rasmussen and Williams, 2006), but are known to have trouble with discontinuities in surfaces. That said, we have no evidence to suggest there are discontinuities in our likelihood surfaces, and indeed the smoothness of most likelihood surfaces is one of the reasons we emulate likelihoods / summary statistics rather than raw model output.

Overall, we agree that further developments should be investigated in this area of research. However, despite the differences, we believe it is encouraging to see emulator posteriors do not exclude the parameter space that the bruteforce suggests, and often agree well with bruteforce posteriors.

**RC 2 - Overall I miss quantitative statistical information about the fitness (model performance) stemming from the parameters obtained via the emulator and the "bruteforce" method against (1) synthetic data and (2) observations (e.g. Nash Sutcfliffe or the Kling Gupta Efficiency). This should also be illustrated by scatter plots and figures that show not only the subdaily but also the seasonal cycle in synthetic/real-world data against models.**

**Knowing the time it takes for the calculations to get done is indeed of technical relevance. But here the most relevant aspects (at least in the perspective of BGD) are centered on how the different model realizations stemming from the emulator approach against the traditional approach change the retrieval of optimal parameters (and posterior uncertainties) and in the eddy covariance flux predictions**

**(for which many relevant information is mostly found in supplements). These are especially important to understand the limitations and caveats of the current proof-of-concept exercise (evaluation of the synthetic exercise).**

AC - We thank the reviewer for pointing out NSE and KGE statistics used in the hydrology literature. We agree with the reviewer that reporting quantitative statistical information about model performance is important. Indeed, we provide performance metrics in the paper currently (please see Table 2 in the manuscript). We report RMSE values, which is related to NSE. In addition, we also report deviance values which takes into account the chosen likelihood and are a more relevant approach regarding the Bayesian framework (Hooten and Hobbs, 2015 as cited in the manuscript). For emulator approach, we report both these metrics against synthetic and real-world data. For bruteforce approach, we only report these metrics against real-world data to provide comparison to the emulator performance. As we already know the true values for synthetic data, we did not feel the need to run bruteforce approach to evaluate emulator performance there. But our workflow is ready to do that in case requested.

We did not present predicted vs observed scatter plots for two reasons: 1) The temporal trends are not visible from such plots, therefore, we decided to go with the more informative unsmoothed time-series plots of both predictions and observations. 2) Such scatter plots are easier to visualize when data is plotted against single model run, while the Bayesian approach produces an ensemble probability distribution of runs. Model ensemble means or medians could be used, but, we wanted to provide the CIs, incorporating the posterior parameter uncertainty. Concentration ellipses could also be used, such as in the an example below (Fig 3). However, we still think temporal trends would be missed by such plots, and including both plots in the main text that essentially test for the same thing would be unnecessary. That is also the reason why we provided diurnal cycle plots in the supplementary, while there is the whole time-series plot in the main text.

Seasonal plots of the model ensembles could also easily be made, but seasonal plots



of the data are not possible without gapfilling the data using some other statistical or mechanistic model, which would then result in a model-model comparison rather than a model-data comparison.

**RC 2 - Another missing important aspect is to understand how the overall results change when contaminating the synthetic dataset with noise (with the same characteristics such as the real observations).**

AC - We thank the reviewer for pointing this out. While it is true that our original analysis did not contaminate the synthetic dataset with noise, our synthetic data had certain characteristics as the real observations. Namely, it had the same gaps as the filtered flux data, and the same coarser time-step and small sample size (n=39) as the real SoilResp data.

We agree that testing against a noisy synthetic dataset is an important test. We contaminated the synthetic dataset with noise based on the uncertainties in the fits to field data and repeated the experiment. Emulator approach showed similar performance with the contaminated synthetic data. Please see Fig 1 below. We believe this is a more proper test than our current version. Therefore, we changed the synthetic data experiment in the main text with the one against the contaminated synthetic data.

> L335: We generated a random parameter set for the SIPNET parameters shown in Fig. 2, and ran the model forward with these values (Table S3). In order to give the synthetic data real characteristics, model outputs were reformatted to have the same gaps, time-steps and sample sizes as the data used in this study. Then, the likelihood parameters were calculated from the synthetic dataset, and next, further noise was added by drawing values from their respective likelihood functions to obtain the final synthetic dataset.

**RC 2- It is not analyzed how does the emulator performance changes by the**

**inclusion of more or less data streams.**

AC - While we agree that effect of including more or less data streams in model calibration is an interesting question and an active area of research, we find it to be a more general data assimilation question rather than being emulator specific, and out of the scope of this particular study. We cite two papers that are looking into this question (MacBean et al., 2016; Cameron et al., in prep.). We designed our framework to make assimilation of multiple data streams possible. With more or less data streams, calibration performance of the emulator should still be proportional to bruteforce rather than showing large independent emulator-specific differences.

**RC 2 - There are a few uninformative visuals, like Figure 4 top 2 panels; Figure 3, the som_resp_rate; that could be replaced by more informative elements (new figures, or tables).**

AC - We thank the reviewer for the suggestions. We now replaced Figure 3, som_resp_rate panel with a more informative x-axis range, please see Fig 1 (with contaminated synthetic data) below.

We also agree that top 2 panels of Figure 4 in the manuscript are "busy". However, we have found the current unsmoothed time-series plot to be more informative than smoothed ones or predicted and observed scatter plots with a 1-1 line as it shows the overall temporal trend without biases as mentioned above. The smoothed figures could be drawn only with a gapfilled flux data. However, as it was the unfilled data that the model was calibrated against, this causes some data points to fall out of the the calibrated model CIs as an artifact of gapfilling and smoothing.

Below, we provide both a concentration ellipses version and a smoothed time-series version of model-data comparison (Fig 3 and 4 respectively). We propose to include the smoothed time-series version in the main text, and include the unsmoothed version and the version with ellipses in the supplement for the interested readers, unless the editor suggests otherwise.



**RC 2 - Related to Equation 3, please see the analysis and discussion in Lasslop et al 2008.**

AC - We thank the reviewer for pointing out this paper. Treatment for the asymmetric heteroskedasticity of the flux data is critical for parameter estimation. While Lasslop et al. (2008) argue that the double exponential distribution of fluxes is largely due to a superposition of Gaussian distributions, they showed that distributions of all error estimates still have a Laplacian distribution. It is also assumed that random errors on eddy covariance data would be approximately normal when integrated over a day (Richardson et al., 2010), but in this study we assimilate all fluxes at the half-hourly scale. Therefore, we think the asymmetric heteroskedastic Laplacian distribution choice in our study is justified. We have added a reference to Lasslop et al. (2008).

> L248-250: The error distribution of flux data are known to be both heteroskedastic, with variance increasing with the magnitude of the flux, and to have a double exponential distribution (Richardson et al., 2006; Lasslop et al., 2008). In previous studies, the error distributions of high flux magnitudes and fluxes averaged over time were modeled as Gaussian (Lasslop et al., 2008; Richardson et al., 2010). However, as we assimilate all flux magnitudes at half-hourly time-step we modeled the error distributions of NEE and LE fluxes as asymmetric heteroskedastic Laplacian distribution.

**RC 2 - The argumentation behind the sufficient statistics is not sustained by the experiment.**

AC - We thank the reviewer for pointing this out. We will now extend the following section in the discussion. Please also see our next response.

> L493-496: A second addition to our scheme was that we included a further generalization of emulation of the sufficient statistics (T) surface. T is,

by definition, sufficient to estimate the simulator (process model) parameters in the MCMC. Unlike emulating the likelihood (this study, Oakley and Youngman, 2017; Kandasamy, Schneider and Poczos, 2015) or the posteriors (Gong and Duan, 2017), emulating T allows us to estimate parameters that are not part of the process model but are part of the statistical data model (the likelihood) as well. In this study, we tested the sufficient statistics emulation for the SoilResp data and updated Gaussian likelihood precision parameter in the MCMC together with other process model parameters. This residual parameter includes both data error and model structural error, and it is not possible to distinguish one from the other with this approach (Van Oijen, 2017). However, when we apply the same calibration scheme to different process models at the same site, because the observation error in the data are the same, the difference in the posteriors of this residual parameter (Fig. S2) could give us clues about the model structural errors of models relative to each other, as we demonstrate in this study as a proof-of-concept. However, in our study use of multiplicative bias parameter further obscures the difference between observation and model structural error.

**(this is a comment by R1 included here for completeness) RC 1 - Published methods for Bayesian calibration increasingly take into account that models are imperfect. There is a discrepancy between model output and reality, even at the best possible setting of model parameter values. This discrepancy is often modeled as a Gaussian Process for which - in the Bayesian calibration - the hyperparameters are estimated together with the regular model parameters. Likelihood-emulation precludes including discrepancy-estimation because model outputs are not calculated during the MCMC. Please add a discussion of this limitation of your approach.**

AC - This is an important point. First, it is worth noting that our current scheme does allow the inclusion of this discrepancy in terms of a bias and variance terms that are



estimated together with the regular model parameters. Indeed, the ability to fit the variance term at the same time as the parameters is precisely why we switched to emulating summary statistics. That said, it is true that in the current implementation the bias term is assumed to be a fixed constant, not varying dynamically, and the soil respiration variance is assumed to be homoskedastic (though the tower fluxes are not). Second, we would argue that our approach does not preclude a more flexible bias specification. Indeed, while beyond the scope of the current paper, conceptually it should be possible to use a bias-variance decomposition to separate our single emulator of the error surface into two separate emulators for bias and variance terms. Similarly, our approach does not preclude specifying a likelihood with a temporally autocorrelated error (which is functionally equivalent to a GP error model in the time dimension), and augmenting the emulator with the autocorrelation parameter similar to how we augmented the emulator with the bias term. As discussed in the paper we instead chose to approximate this as an effective sample size correction, both for computational efficiency and because accounting for autocorrelation in an asymmetric heteroskedastic Laplace is more complicated than doing so in a multivariate Normal. Finally, in PEcAn, we are working towards a more general framework for model-data integration that takes into account initial condition / driver / parameter / model structural uncertainty in calibration and prediction. However, this is still work in progress.

We will include the following in the discussion, at the end of section 4.2 after the paragraph in the response above:

L497: Indeed, implementation of a more formal way of accounting for model structural error (also called the discrepancy between model output and reality) in our emulator scheme is one of our planned next steps. Explicitly specifying a model discrepancy term and estimating it through MCMC would allow us to account for all sources of model predictive uncertainty (Van Oijen, 2017). However, determining the expected form of discrepancy in order to learn about model parameters realistically could be dif-

ficult due to lack of mechanistic knowledge of the underlying processes (Brynjarsdottir and O'Hagan, 2014). In that sense, accounting for discrepancy in model calibration is not an emulator approach specific issue. For a novel approach investigating model structural uncertainty through a modular modeling framework see Walker et al. (2018), which could be useful for modeling prior knowledge about discrepancy in ecosystem models in the future. Because of the unknowns about the discrepancy functions, it is common to use Gaussian processes to model the discrepancy (Kennedy and O'Hagan, 2001). Even then, only with realistic prior constraints about the process, calibrated model predictions will be unbiased (Brynjarsdottir and O'Hagan, 2014). For an example of addressing discrepancy in calibration that combines likelihood-emulation approach with importance sampling, see Oakley and Youngman (2017) where they inflated simulator uncertainty to account for simulator discrepancy instead of explicitly specifying a prior for it, in order to make the likelihood tractable. When likelihood function becomes intractable, techniques using likelihood-free inference could also be a remedy (Gutmann and Corander, 2016).

**Cited references (that were not mentioned in previous correspondence or in the manuscript):**

Brynjarsdottir and O'Hagan, 2014, Learning about physical parameters: the importance of model discrepancy, 30, IOP, doi:10.1088/0266-5611/30/11/114007

Dietze et al., Beyond MCMC: Data-constraint and error propagation in a dynamic terrestrial biosphere model through Bayesian model emulation, American Geophysical Union, Fall Meeting 2009, http://adsabs.harvard.edu/abs/2009AGUFM.B44A..02D

Erickson, C.B., Ankenman, B.E., Sanchez, S.M., 2018, Comparison of Gaussian process modeling software, European Journal of Operational Research, 266(1), 179-192 https://doi.org/10.1016/j.ejor.2017.10.002



Fisher, R., A., On the mathematical foundations of theoretical statistics, Philosophical Transactions Of The Royal Society A.222:309-368, doi:10.1098/rsta.1922.0009

Gupta, H. V., M. P. Clark, J. A. Vrugt, G. Abramowitz, and M. Ye (2012), Towards a comprehensive assessment of model structural adequacy, Water Resour. Res., 48, W08301, doi:10.1029/2011WR011044.

Loeppky, J. L., Sacks, J., Welch, W. J. (2009). Choosing the sample size of a computer experiment: A practical guide. Technometrics, 51(4), 366–376. doi:10.1198/TECH.2009.08040.

Mikusheva, A., Lecture 4 Sufficient Statistics, course materials for 14.381 Statistical Methods in Economics, Fall 2011. MIT OpenCourseWare (http://ocw.mit.edu), Massachusetts Institute of Technology. Last access on 14/05/2018.

Rasmussen, C. E. (1996). Evaluation of Gaussian Processes and Other Methods for Non-linear Regression. PhD thesis, Dept. of Computer Science, University of Toronto. http://www.kyb.mpg.de/publications/pss/ps2304.ps

Walker, A. P., Ye, M., Lu, D., De Kauwe, M. G., Gu, L., Medlyn, B. E., Rogers, A., and Serbin, S. P.: The Multi-Assumption Architecture and Testbed (MAAT v1.0): Code for ensembles with dynamic model structure including a unified model of leaf-scale C3 photosynthesis, Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2018-71, in review, 2018.

---

Interactive comment on Biogeosciences Discuss., https://doi.org/10.5194/bg-2018-96, 2018.

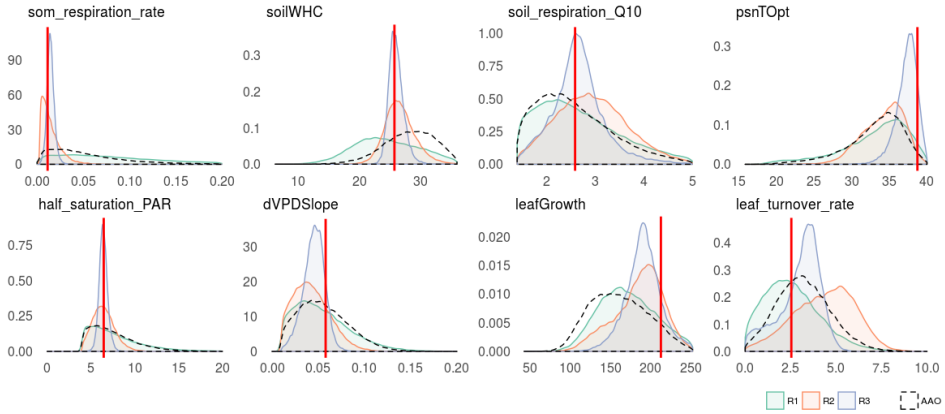


**Fig. 1.** New Figure 3, contaminated synthetic data version.



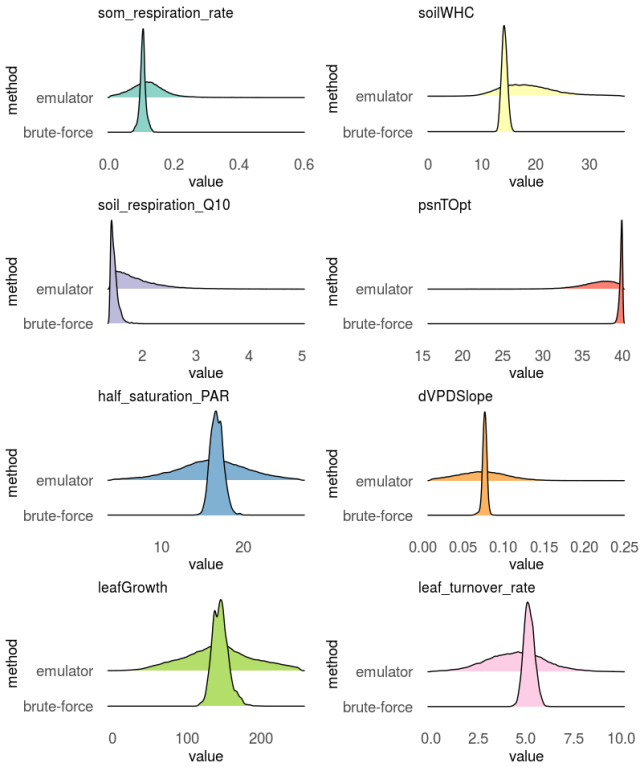


**Fig. 2.** New Figure 5, after sampling joint posterior distribution instead of marginal distributions in between iterative emulator rounds.

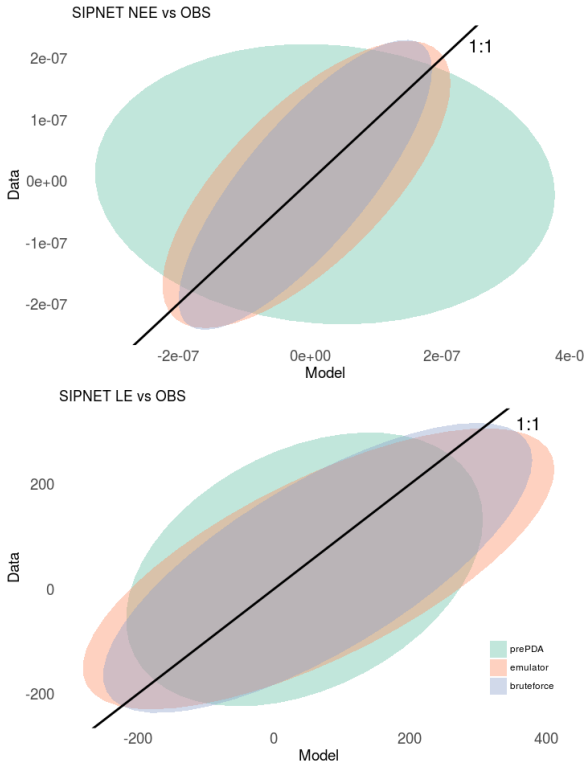


**Fig. 3.** Predicted vs observed, concentration ellipses version.



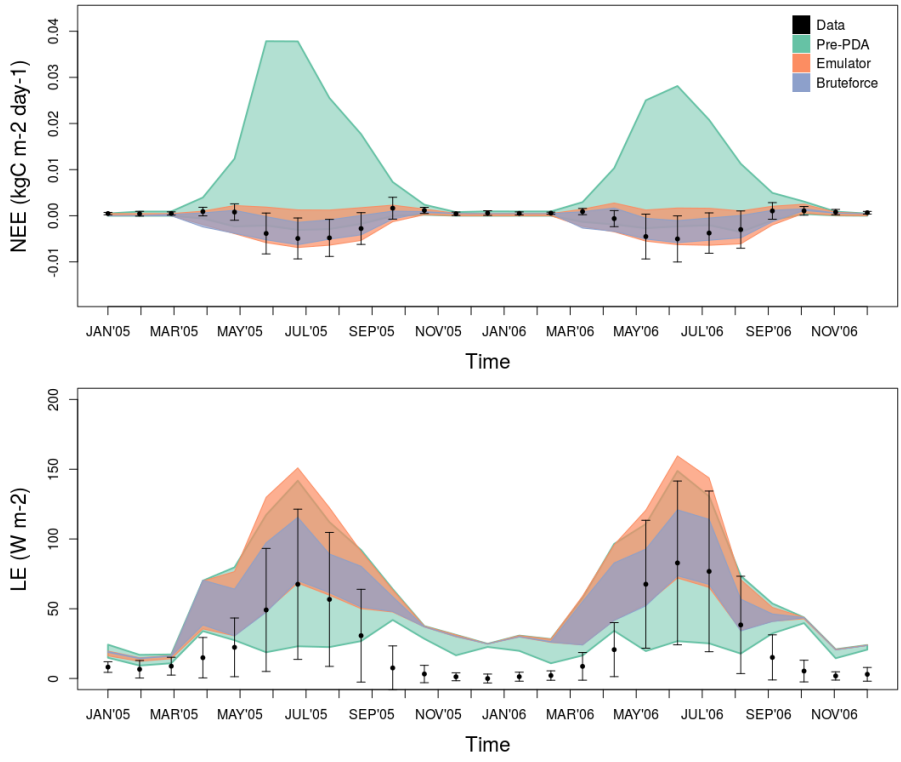


**Fig. 4.** Predicted vs observed, smoothed time-series version.