Machine Learning for a-posteriori model-observed data fusion to enhance predictive value of ESM output

Claudia Tebaldi (JGCRI), Z. Jason Hou (PNNL), Abigail Snyder (JGCRI), Kalyn Dorheim (JGCRI)

Focal Areas: The proposed research is consistent with focal areas 2 and 3: when successful, this effort will enhance the predictive value of Earth System Model (ESM) output by identifying and quantifying its systematic deviations from available observations and by correcting and recalibrating to existing data. The product is a hybrid process & data-driven modeling tool whose estimation can also bring insights through pattern discovery recognition on model deficiencies on the one hand and data gaps on the other.

Science Challenge: Predictability of the Earth System is a challenge that needs to be met by progress in process-based modeling. While process-based modeling evolves, there remains the need to correct and recalibrate model output in order to enhance its predictive value for decision-makers and for impactful research and modeling. There exist specific challenges to this correction when the target is as complex, multi-dimensional and century-scale long a process as the output of an ESM. The computational tools needed to systematically compare model outputs to observations, and then develop and integrate an error model for ESM model correction, are complex and costly. These computational efforts would greatly benefit from artificial intelligence and machine learning (AI/ML) approaches deployed on super-computers.

Rationale: In simpler incarnations, bias-correction has been applied for decades now to both weather and climate models' output in order to ameliorate their representation of the systems they seek to predict. Existing methods however are falling short of addressing emerging needs like joint variable correction (in order to correctly characterize compound effects of climate impact-drivers) and extreme behavior (which we know drives the most significant and costly impacts). They are also often based on simple offsetting (additive or multiplicative) and linear interpolations when addressing the need for finer resolutions. With our proposed research we recommend advancing bias-correction to a much more comprehensive and effective level than the current practice, enhancing the saliency of model output for decision making and impact modeling with significantly improved predictive power. We want to exploit the richness of observational data products that are accumulating (also thanks to DOE efforts), the power of available and next-generation computing resources, and the DOE investment in Earth System analysis and modeling. The aim is to estimate spatio-temporal, multi-dimensional error models that exploit and respect process representation and understanding as much as possible. These error models are by construction complex and computationally challenging to obtain, but would constitute a leap forward in our ability to correct, calibrate and characterize uncertainties in ESM output. Particular focus would be directed towards correction and calibration of multi-variate output from E3SM and other ESMs, and on the correction and calibration of compound (joint) extremes, with special attention towards those influencing the accuracy and uncertainty of water cycle prediction.

Narrative: ESMs are becoming ever-more complex to increase their fidelity to the real system that they seek to emulate. Efforts to bring observational data's insights into their parameterization and calibration have recently started to be successfully carried out with the use of AI/ML, and so have been methods for data assimilation. Increasingly more sophisticated deployment of AI/ML tools within these domains, which we here label "a priori" modeling enhancements, are sure to bring further progress in modeling fidelity and reliability. As this happens, an "a-posteriori" correction and recalibration of ESM output still remains an important step in further enhancing model performance and predictive skills. This is true both of weather prediction models/ensembles and of climate/earth system models output. The proposed research positions itself within this "a-posteriori" model improvement research domain. We define a

DOI: https://doi.org/10.2172/1769740
DOE BER, AI4ESP 2021 #AI4ESP1131 (https://www.ai4esp.org)

strategy that will provide sophisticated and comprehensive solutions to what could be for simplicity labeled by the familiar term of bias-correction by harnessing AI/ML tools.

The DOE research environment is a prime domain for the success of such effort as it houses all of the pieces needed, including advanced computing alongside cutting edge Earth System Modeling, as well as extensive observational products collection and curation. These state-of-the-art capabilities are a critical ingredient to a transformative solution to the problem of bias correction.
The research effort envisioned here consists of:

a) Building *enriched observational products* by merging existing datasets in space and time, and deriving the best and longest multi-dimensional representation of the variables of interest. This would be done by using ML tools to fuse and infill disparate sources of observations and obtain a regular (in time and space) multi-dimensional field of high-frequency quantities. In this step, the problem of determining the optimal spatial resolution at which to perform the correction (see next steps b and c) would also be explicitly tackled. That is, the level of spatial disaggregation that wins the trade-off between the observational product field resolution and the model's field resolution preserving the highest value of information would be determined.
b) Focusing on the *joint correction of multiple variables* in order to preserve their relation, important in projecting ahead corrected realizations of variables that contribute interactively to impacts; this is arguably especially true when considering water cycle-relevant outcomes, where precipitation, temperature, humidity, and winds can converge to create a hazard. This would require fitting complex, multidimensional error models from the comparison of ESM output to observations, and then applying the error model to the future/out-of-sample output stream. ML/AI techniques would be key in handling the large data sets involved, and in learning the complex, non-stationary spatio-temporal relationships among them.
c) Paying special attention to the correction and preservation of compound *extremes*, again the most consequential yet less frequent hazards. This will likely require fitting error models that are specific to different quantiles of the marginal and – importantly – joint distribution of variables, focusing on the lowest and highest quantiles. The convergence of ML tools (e.g., convolutional and recurrent neural networks, ensemble random forest and gradient boosting machine, or extreme learning machines for learning mixed types of extremes data) and statistical tools (e.g., extreme value theory and copula for representing compound extremes) here should be key to tackle a problem that, at least when using observations, will likely be marred by low statistical power (few observations of – particularly – joint extremes) and where statistics of extreme values should be still relevant to make the best of the scarce observations, even in the age of big data. The aim is to optimize the hybridization of statistical tools that are designed to unearth and represent dependencies in data-sparse regions (extreme value statistics is developed for exactly these problems, and much progress has been made on the estimation of tail dependency from sparse data in the last years) with unsupervised learning techniques that could take those structures and transfer/extend them under the assumptions that they remain valid for other spatiotemporal domains where data is altogether lacking.

ML techniques will be employed in order to handle the big data nature in the bias-correction problem, and will be built upon multiple observational datasets at spatial and temporal scales with potentially significant mismatches, as well as on integrated large ensemble sets of climate model output.

The approach will consist of both computationally and scientifically novel efforts in

1. Spatiotemporal pattern recognition for determining the multi-dimensional spatiotemporal process that we would call "error", which will imply interpolation and data filling (when observational products do not have perfect coverage in time and space but are still considered our best source of

the true process characteristics), and differencing of the resulting fields from the corresponding model output (which we call *in-sample*);

2. Structural cross-dependence and joint distribution analyses of this spatially and temporally varying error process and identification of robust features in high-dimensional space and time;
3. Development of predictive understanding of the errors and application of the resulting error models to the *out of sample* model output, i.e., the parts of the simulations that do not overlap with observations, but for which the assumption is made that the same error characteristics would be found.

Physical knowledge and statistical estimation of dependency structures about the errors will be used to guide the ML development by providing physical explanations, or priors of the joint distributions, or structural priors for the ML architect. Ensemble ML approaches (e.g., random forest, gradient boosting, support vector regression, recurrent/convolutional neural networks) will be integrated to look into error data in the forms of ensemble, imagery, and time series. The models will be validated, optimized, and integrated to enable accurate error characterization, prediction, and reduction, with quantified uncertainty. The ML-based predictive understanding of error distributions with respect to spatial locations, temporal scales, and physical factors can in turn identify systemic biases or structural errors for correcting physics-based priors or guiding model development.

We propose that the first target of such effort should be E3SM output, to enhance its predictive value for stake-holder and decision-makers. However, a wider use of the results would be the uptake by the climate impact research community, which also requires bias-corrected climate information. For example, any impact analysis that hinges on characterizing exceedances of thresholds, defined as absolute rather than relative values, needs bias corrected climate information. Much of the currently accepted (and expected) options in the field of global Multi-Sector Dynamics models come from the Inter-Sectoral Impact Model Intercomparison Project (ISIMIP) project[1]. While the ISIMIP project does an excellent job of documenting their methods, in-house adoption can be challenging, given the inherent constraints following from ISIMIP choices in terms of models/scenarios/number of ensemble members available. Further, the ISIMIP method, like most bias correction/statistical downscaling methods in the literature, still relies on classical statistics (most are simple regression fitted between model output and observable quantities), which can be challenged by the quantity of data and nonlinear processes that must be treated across space, time, and multiple variables (all that multiplied by numerous models and scenarios). Our results would constitute a substantial improvement over the existing products that the impact research community uses. Lastly, we would expect the error models' results to be valuable in identifying areas for high potential or strong need of improvement from the modeling perspective and/or areas in need of more focused observational efforts, thus extending the value of this research towards model "a priori" improvement (to use the label that we introduced earlier).

The algorithms will be easily transferable and generalizable to be efficiently deployable for all major types of climate model outputs of interest, by developing and testing them on a range of CMIP-class models (after and besides E3SM), for different scenarios and for a range of output variables, at frequency from daily to monthly, seasonal and annual time scales, and for the whole length of ESM simulations (relying on the assumption that the error model can be applied to the future simulation periods, which is the unavoidable assumption of bias correction). The implicitly developed capabilities of infilling and spatial downscaling would be also made transferable and generalizable. The developed open source algorithms with documentation would facilitate the use of the approach and tool by the community, and FAIR (findable, accessible, interoperable, and reusable) data practices would be applied to the intermediate and final data products.

---

[1] https://www.isimip.org/gettingstarted/input-data-bias-correction/