Machine Learning for Non-parametric Data Assimilation in Hydrological Models

Poster category: Science

Dr. Craig Pelissier, code 606
Advanced Software Technology Group
Goddard SPACE FLIGHT CENTER

Jonathan Frame
University of Alabama
Department of Geological Sciences

Abstract:

We test a method for hydrologic modeling predictions of soil moisture with a hybrid machine learning (ML) + physics-based modeling approach. This method is an alternative to data assimilation, addresses a grand challenge of integrating machine learning with physics, and has an added benefit in that it makes dynamic corrections to model structural error. Dr. Pelissier has developed a parallelized machine learning code for Gaussian Process Regression (GPR) which is used for the ML component, and we use the Noah-Multiparameter (Noah-MP) land surface model as the physics-based component of this hybrid. We test this method over annual soil moisture cycles at FluxNet towers with high quality observations. The results show that this hybrid approach significantly improves the out-of-sample soil moisture predictions as compared to those made by a calibrated Noah-MP model. We also compare the GPR with 'traditional' data assimilation. We ran the Noah-MP model with an Ensemble Kalman Filter (EnKF). The results show a similar performance improvement between GPR and EnKF when run in sample, but EnKF provides no benefit when making predictions after only a few time steps following an observation. Our results show that this hybrid approach continues improving model predictions even without soil moisture observations. This has significance for improving the efficiency of satellite data assimilation into large scale hydrologic models.

Background and Introduction:

Machine Learning has been a part of Hydrological modeling since at least the early 1990s [1-3], but is emerging as an important method to make predictions and understand the hydrologic cycle [4]. A major concern is that data driven predictions lack the adaptability of physically based models. Our approach leverages the complementary aspects both of physically based models and machine learning, which **is absolutely critical for the future of hydrology.** Kratzert et al. [5] showed that machine learning can produce, on average, better streamflow predictions in *ungauged* basins than traditional Hydrology models can produce in *gauged* basins. To achieve this, Kratzert et al. used a set of catchment attributes derived from a combination of remote sensing, soils maps, and climate model output (known as the 'CAMELS' dataset; [6]) as inputs to allow his ML algorithm to differentiate between different types of catchment behaviors. This result means that there is sufficient information in the remote sensing and climatic data record to differentiate between at least a significant portion of diverse catchment-specific rainfall-runoff behaviors.

Data Assimilation
Data assimilation (DA) is the most common method for using satellite data to improve large-scale hydrological models like NASA-LIS [7], but has several limitations:

- DA only updates the model state (or state and parameters [8]), and does not help mitigate model structural errors or errors due to missing inputs.
- DA requires estimates of uncertainty distributions over the model and observations, which can be difficult to estimate, can change over time and in different conditions, and if mis-specified can lead to significant information loss [9].
- DA is typically only useful for mitigating random errors, and the assimilated observation data must be mapped to the model's climatology [10].
- Kumar et al. [11] identified major unmodeled process in the current generation of hydrological models contained in NASA-LIS, and noted that assimilation of data associated with unmodeled processes can produce spurious results. These limitations apply even to modern data assimilation techniques [12].

We propose that ML provides a direct approach to merge data with models to account for model structural uncertainty, model bias, missing inputs, and unmodeled processes.

Methods/Approach:

Dr. Craig Pelissier at the NASA Center for Climate Simulation (NCCS) has developed code for high performance Gaussian Process Regression (GPR). A graphical representation of a Gaussian process is shown in Figure 0. We have applied this GPR code with the Noah-MP hydrology model. The GPR is trained to predict the difference between the hydrology model and soil moisture observations using the model states and atmospheric conditions all measured at FluxNet tower sites (Figures 1 & 2). **This is a ML-based data assimilation scheme that can update model structure** (not just the model's state).

GPR assumes a zero mean prior. This is an important characteristic which makes it suitable for this application the prediction defaults to the physical model (in our case Noah-MP) and updated with a contribution from the GPR only if it adds value. This is important as hydrologic conditions often fall outside of the historical record, i.e., major floods and droughts.

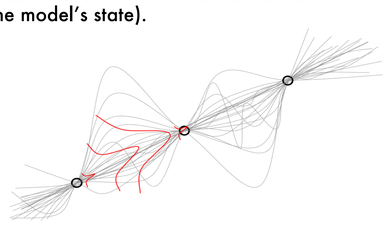

Figure 0. A graphical representation of a one dimensional Gaussian Process. This figure shows an ensemble of functions that pass through observations, with lower variance closer to the observations.

The GPR prediction is merged with Noah-MP in two distinct ways. Post-processing is an application of ML to predict differences between a time series of dynamic model simulations and time series of observations. Post processing allows us to extract that information from observations to improve model predictions, but it doesn't tell us what is wrong with our dynamical systems models. To address this harder problem of learning deficiencies in the original model we apply the GPR prediction dynamically to the model state at each time step.

The results are updated model-simulated variables that better match the future observation data, assuming that the GPR has identified trends in the model error, perhaps due to missing model inputs or unmodeled processes.

A typical downside of GPR for machine learning is that they usually take a long time to train. To overcome this obstacle our code uses Sparse Pseudo Input GPR [14], which speeds up training by orders of magnitude. We run the GPR code on the NCCS Discover, and training time is usually just a matter of seconds.

Project Data for Land Surface Modeling Experiments
We evaluated predictive performance of the approach described for Soil Moisture. Model input data was aggregated from NLDAS. Model performance was assessed against observations at 10 FluxNet Towers with consistent data sets.

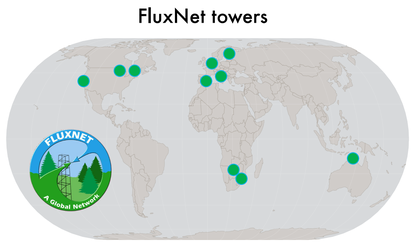

Figure 1 (above). Location of Ten FluxNet tower sites

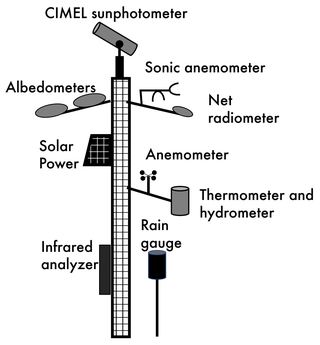

Figure 2 (right). Schematic of a FlexNet Tower with atmospheric and hydrologic measurement devices

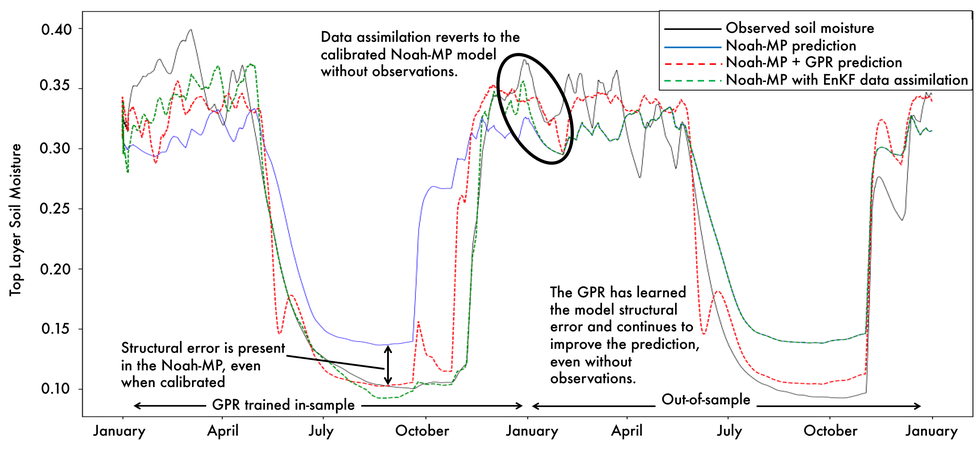

Figure 3. Results using the GPR as a post processor at the Blodgett (Ca) FluxNet tower site with a direct comparison to the Ensemble Kalman Filter. The data shown above has been smoothed (two day running average) for clarity. The first year in this graph is "in-sample", meaning the GPR was trained during this time, and the Ensemble Kalman Filter functions as intended.

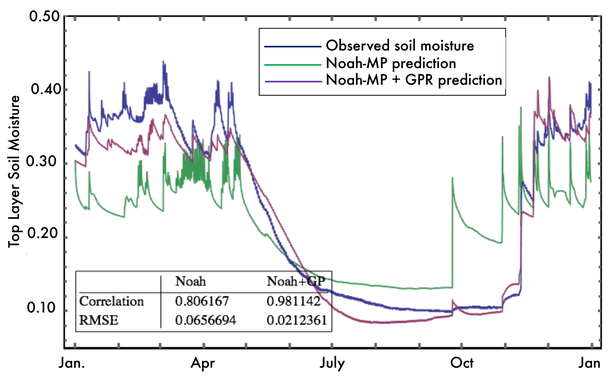

Figure 4. Results using the GPR to update the model state dynamically. These results are out-of-sample, meaning that the GPR was trained and tested on separate data sets. This implies that the GPR has sufficiently learned model structural error to improve soil moisture predictions when no observations are available.

Discussion and conclusions:

Our results demonstrate a model performance improvement with the GPR for both post processing (Figure 3) and dynamic state updating (Figure 4). These results are a proof of concept, that we can use machine learning for non-parametric data assimilation in hydrological models. This is a significant step forward for hydrological prediction, and for the integration of machine learning in dynamical systems modeling.

This work has significance to gain a deeper understanding of hydrologic processes in general. We see that the GPR 'learns' the difference between soil moisture observation and the predictions made by an operational land surface model. This implies that there are processes that occur in nature that are not represented in Noah-MP, which in turn implies that there is opportunity to improve our understanding of hydrologic process and the physically-based dynamic model itself.

Further work

- Extend this procure to the rest of the FluxNet tower site and other high quality hydrological data sets, including the CAMELS catchments.
- Using the Machine Learning to make predictions across sites. While it is relatively straight forward to use data collected at a site to make predictions at that site, it would be valuable to be able to make predictions at one site with data collected at a different site. For this to work the GPR will have to learn hydrologic model structural error that is not site specific.
- Use Machine Learning to assimilate remote sending data into land surface model predictions. This is a direct extension of the bullet listed above. Remote sensing data necessarily represents a gridded area of Earth's surface, so it is necessary that the GPR to learn geo-spatially averaged model structural error.
- An additional path for future research is using the dynamic correction factor to study hydrologic nonstationarity, which is a challenge for long term water resources management and hydrological forecasting. Figure 4 shows a map of how the surface water hydrologic response to precipitation is changing across the united states from 1979 – 2004. This future work will *quantify* and *model* nonstationary (*i.e.,* changing) hydrologic behaviors of large scales (continental and global) using a combination of the GPR, process-based land surface modeling with NASA's Land Information System (NASA-LIS), and remote sensing (RS). We can analyze nonstationarities with this hybrid ML + process-based modeling approach.

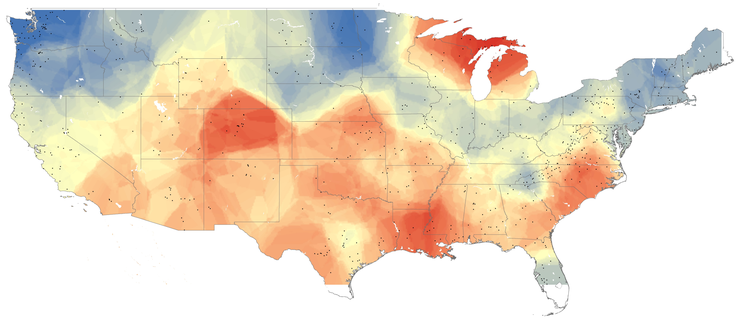

Figure 5. Time series analysis results of hydrologic nonstationarities across the continental United States. Future work will include expanding the capability of the hybrid ML + Process based modeling to predict large scale hydrologic processes assimilating remote sensing data, and will provide a means to analyze hydrologic nonstationarities in greater detail than current time series methods.

References:

References:
1. French et al. (1992). *J. Hydrol.* 137, 1-31. (1992).
2. Smith. M.S. Thesis, West Virginia Univ. (1992)
3. Hsu et al. *Water Resour. Res.* 31, 2517–2530, (1995).
4. Sellars. *Bull. Am. Meteorol. Soc.* 99, ES95–ES98 (2018).
5. Kratzert et al. *Hydrol. Earth Syst. Sci.* 22, 6005–6022 (2018).
6. Addor et al. *Earth Syst. Sci* 21, 5293–5313 (2017).
7. Reichle. Advances in Water Resources, 31(11), 1411–1418. (2008).
8. Moradkhani et al. Adv. Wat Res, 28(2), 135–147. (2005).
9. Nearing et al. Water Resour. Res. 54, 6374–6392 (2018).
10. Kumar et al. Water Resour. Res. 48, 1–16 (2012).
11. Kumar et al. Hydrol. Earth Syst. Sci. Discuss. 12, 5967–6009 (2015).
12. Moradkhani et al. In: Duan Q et al. Handbook of Hydrometeorological Ensemble Forecasting. (2018).
13. Ghahramani & Roweis. Adv. Neural Inf. Process. Syst. 11, 431–437 (1999).
14. Snelson & Ghahramani. Adv. Neural Inf. Process. Syst. 18 1–24 (2009).

Acknowledgements:

Dr. Daniel Duffy
Dr. Grey Nearing
Dr. Christa Peters-Lidard
Dr. Soni yatheendradas
Stewy Slocum
Chris Culver