Surrogate-enabled inference

K. Sargsyan
Sandia National Laboratories, Livermore, CA
September 26, 2018

1 Setup

Consider a model $Z(\boldsymbol{x}, t, \lambda)$ that produces discretized output for a given QoI (in this case, GPP only is considered) Z over a period of time $t \in \Omega_t$ and over space $\boldsymbol{x} = (x, y) \in \Omega_x$ (typically a rectangular region)*, for a given parameter vector $\lambda \in \mathbb{R}^d$.

- Raw data: ensemble of daily simulations is available, i.e. $\{Z(\boldsymbol{x}_i, t_j, \lambda_n)\}$ for $i = 1, \dots, I$, $j = 1, \dots, J$ and $n = 1, \dots, N$, where $I = 61 \times 41$, $J = N_y \times 365$ and $N = 2000$. Number of years is set to $N_y = 24$.

- Monthly averages are taken first, so, with some abuse of notation, let us set $J = N_y \times 12$.

- The actual QoI to be analyzed is the month-by-month average over all years, i.e., for $j = 1, \dots, 12$
$$\bar{Z}_j(\boldsymbol{x}, \lambda) = \frac{1}{N_y} \sum_{k=0}^{N_y - 1} Z(\boldsymbol{x}, t_{j+12k}, \lambda) \tag{1}$$

- We have several sites available, and for the $S = 8$ sites that are in the region Ω_x, we identify the closest $\boldsymbol{x}_i$, see Figure 1. Therefore, the QoIs for which we build surrogates are
$$Q_{ij}(\lambda) = \bar{Z}_j(\boldsymbol{x}_i, \lambda), \text{ for } i = 1, \dots, S \text{ and } j = 1, \dots, 12 \tag{2}$$

2 Surrogates

A total of $12S = 96$ surrogates are build, using BCS and second order Legendre-basis, i.e.
$$Q_{ij}(\lambda) \approx Q_{ij}^c(\lambda), \tag{3}$$
parameterized by basis coefficients c. A subset of inputs lead to 0 GPP - we ignored those inputs (in fact we ignored all below 0.02) for surrogate construction. The surrogates have shown relative errors between $5\% - 15\%$, as demonstrated in two example plots in Figure 2.

The sensitivities for the eight sites over 12 months are shown in Figure 3 (minus the winter months with zero GPP for all ensemble members.).

*where x is longitude, y is latitude.

3 Data

Similarly, we have observational daily data for $S = 8$ sites that is averaged to obtain monthly data before averaging over years to obtain month-by-month data - a total of $12S = 96$ data points $\{h_{ij}\}$ for $i = 1, \ldots, S$ and $j = 1, \ldots 12$.

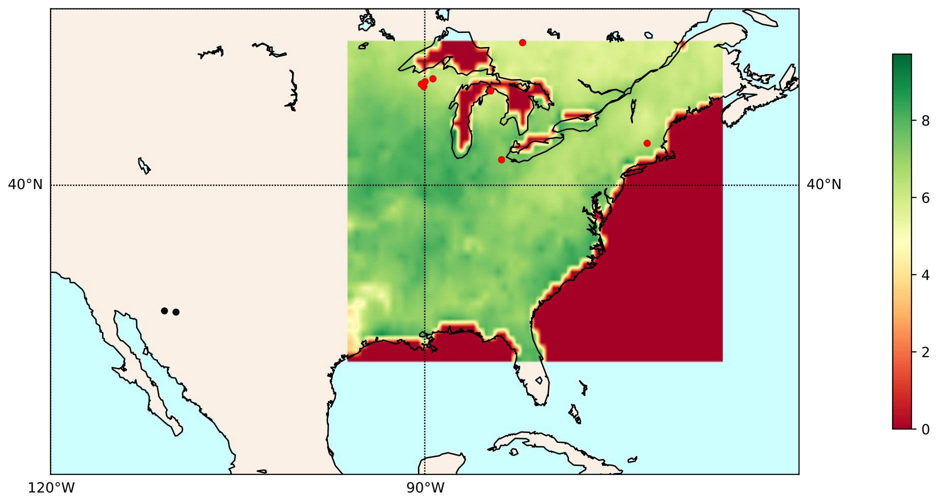


Figure 1: The 8 sites that lie in the region where we have model evaluations. The average GPP field for a month of June (?) is shown.

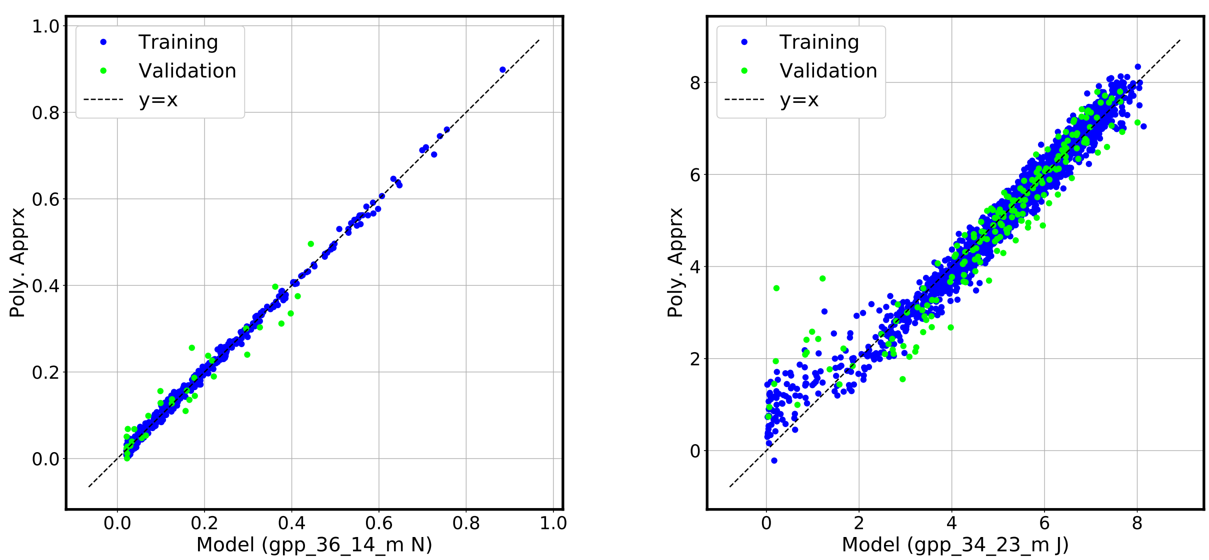


Figure 2: Two representative surrogates.

4 Inference

4.1 Classical

The classical surrogate-enabled inference setup: iid Gaussian noise on $h_{ij} - Q^c_{ij}(\lambda)$. [More details to write later]. We selected the top four most sensitive parameters `gdd_crit, nue_tree, crit_day` for inference. See Figures 4 and 5 for the posterior predictive. [MCMC and parameter posteriors to come later]

4.2 Model error

[to come later as need arises.]

5 Dimensionality reduction via Karhunen-Loève expansions

[ongoing] [will add technical details later]. Plots of mean field and eigenvalue decay are in Figure 6.

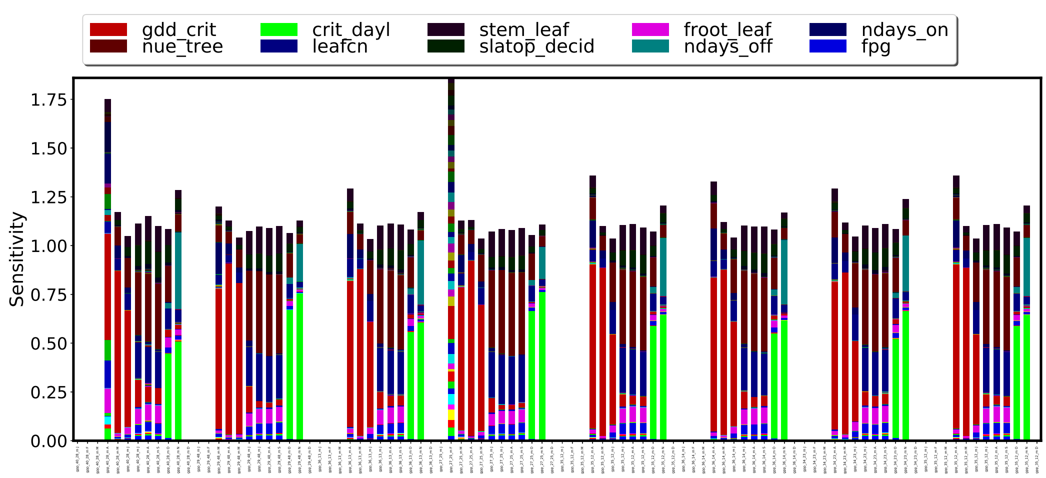


Figure 3: Sensitivities of GPP for 8 sites (the winter sites are not shown due to dominantly zero-ed GPP).

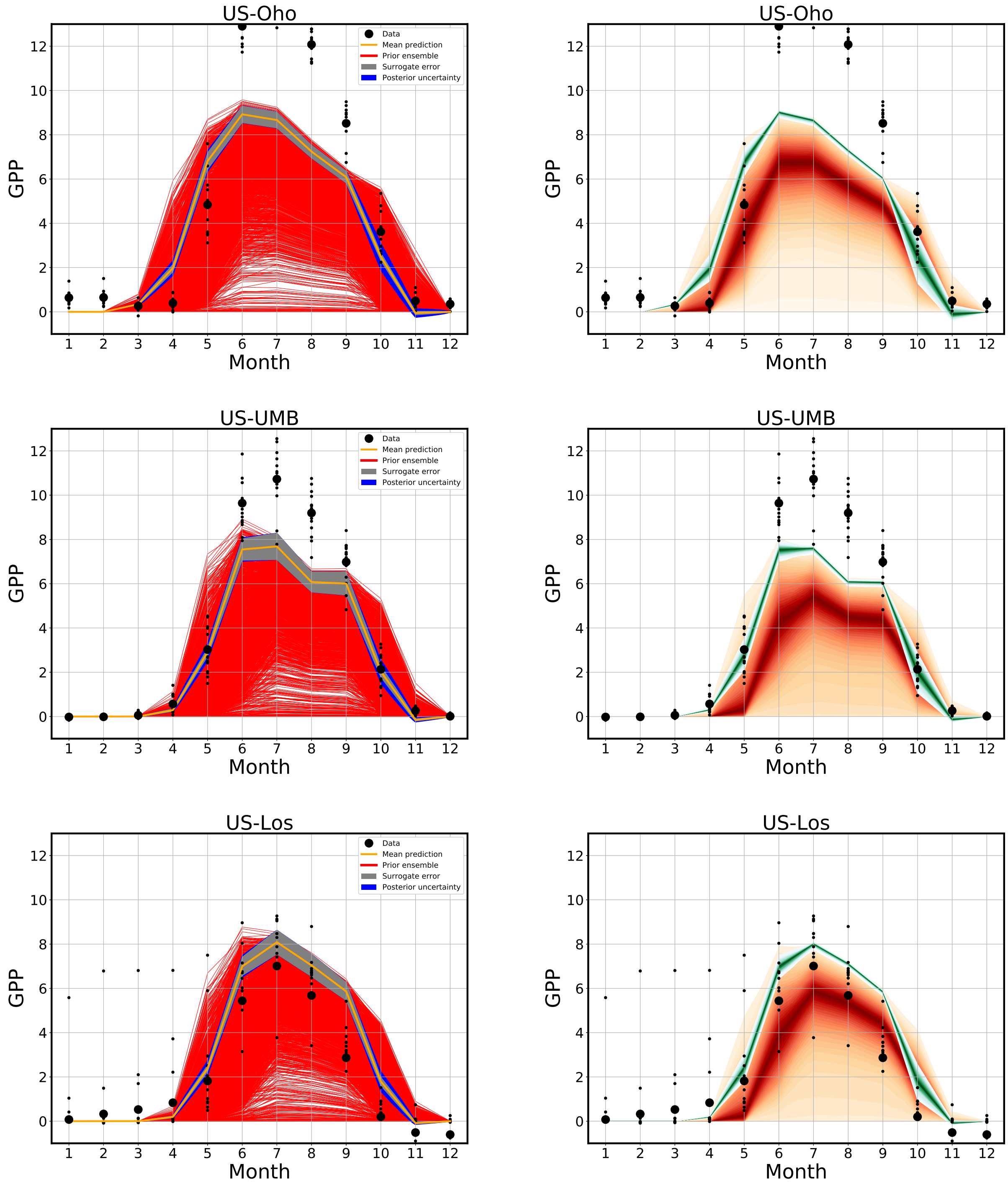
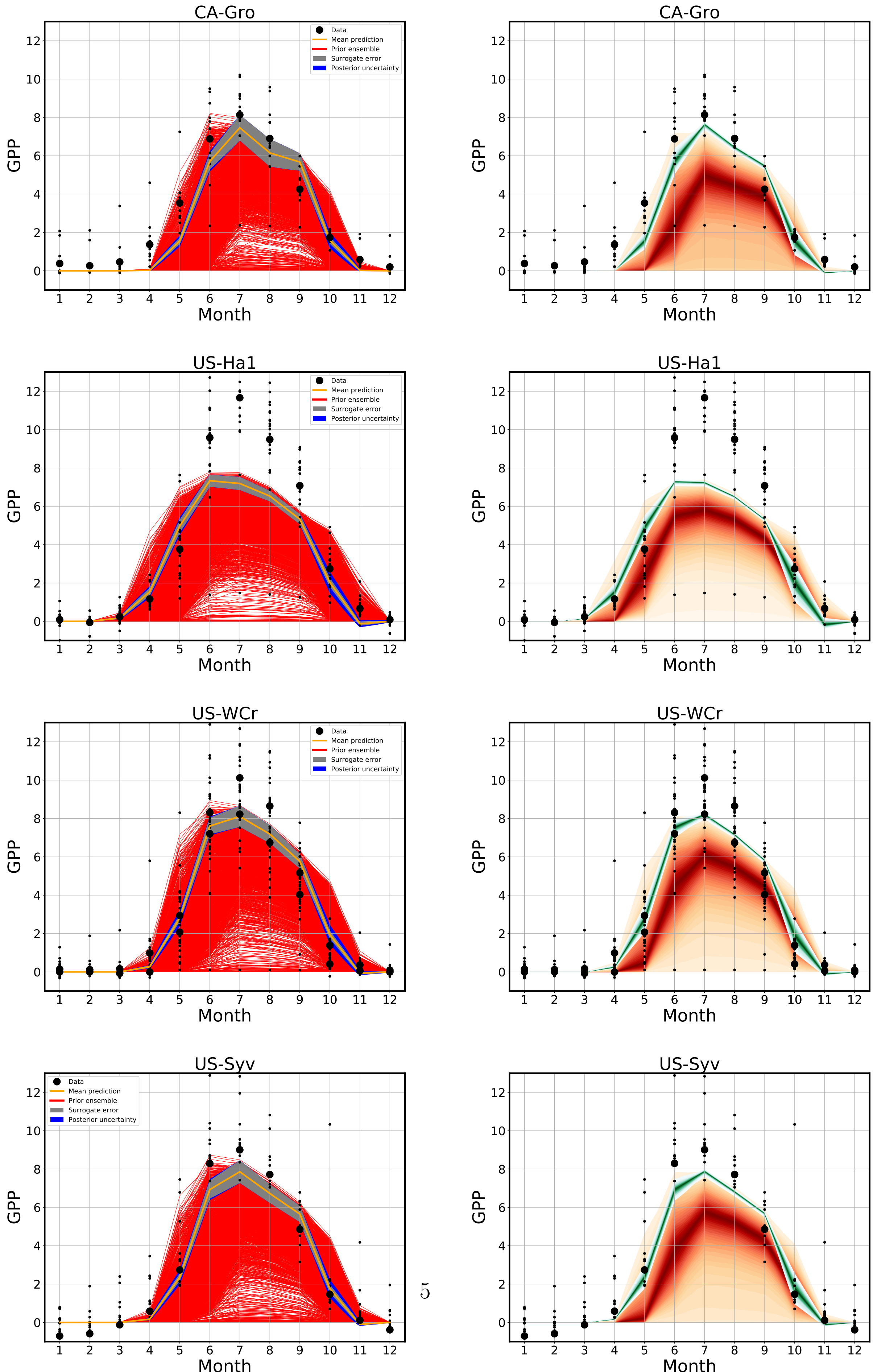


Figure 4: Inference1. Lareg black dots are the mean data points, smaller ones are all samples across years.


CA-Gro
US-Ha1
US-WCr
US-Syv
GPP
Month
Data
Mean prediction
Prior ensemble
Surrogate error
Posterior uncertainty

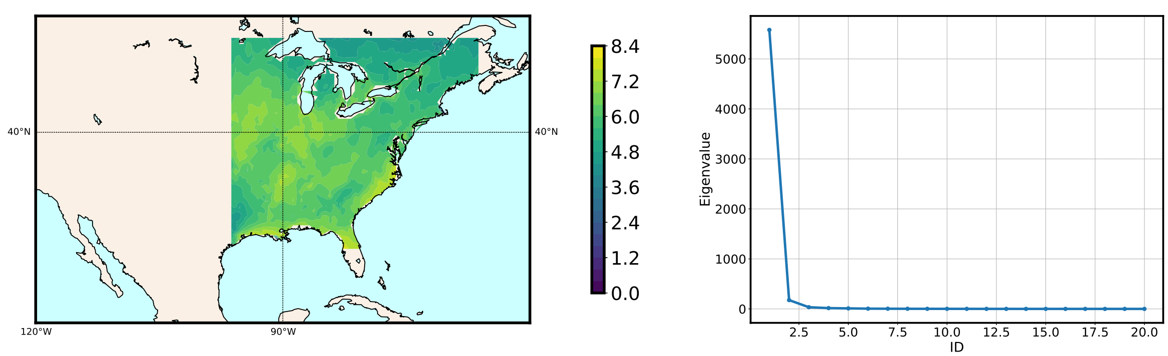


Figure 6: Mean KL field and eigenvalue decay.