Supplementary material

A Vecchia approximation

Here we provide detail and empirical work supporting the discussion in Section 3.3.

A.1 Review

The crux of the Vecchia approximation is a simple identity linking joint and conditional distributions. Consider a vector $\mathbf{Y}_n$ of responses, like one filled with $\bar{y}_i$-values in Section 3.2. The joint probability of $\mathbf{Y}_n$ can be factorized as

$$\begin{aligned} p(\mathbf{Y}_n) &= p(y_1)p(y_2 \mid y_1)p(y_3 \mid y_2, y_1) \dots p(y_n \mid y_1 \dots y_{n-1}) \\ &\approx \prod_{i=i}^{n} p(y_i \mid y_{c(i)}) \quad \text{where} \quad c(i) \subset \{1, \dots, i-1\}, \end{aligned} \tag{1}$$

where the second line is an equality if $|c(i)| = i - 1$ and an approximation otherwise. This relationship is true for all $n!$ reordering of the indices. The quality of the approximation depends on the ordering, $|c(i)|$ and the choice of the subset of $\{1, \dots, i-1\}$ it comprises. Such an approximation is advantageous in particular for GPs, or any MVN-based joint distribution, because the requisite conditionals are available in closed form (Eq. 4 in the main document) and the matrices/decompositions are limited by the size of $c(i)$. Taking $|c(i)| = \min(m, i-1)$, a common simplifying choice, means that each conditional in the product in Eq. (1) requires flops in $\mathcal{O}(m^3)$. If $m \ll n$ this cost represents an enormous computational savings despite that the product involves n of them, i.e., $\mathcal{O}(nm^3)$.

Some additional implementation details make things even faster and allow for distributed computation. We have been satisfied with the performance of the defaults offered by the R referenced in Section 3.3, synthesizing many of those elements including composition of the $c(i)$. We settled on $m = 30$ by entertaining an out-of-sample prediction exercise similar to that reported by Sauer et al. [2023], which showed diminished returns for larger m-values. Perhaps the most important aspect of that software, compared to other methods based on the Vecchia approximation, is that the effect of an input-dependent lengthscale γ can be learned along with scale τ^2 and (scalar) nugget g via Fisher scoring [Guinness, 2021]. The authors call this the "Scaled Vecchia" approximation; however we'll simply refer to it as "Vecchia" with the understanding that input-dependent lengthscales are being learned. With this setup we are able to fit GPs with $n \approx 270{,}000$ as described in Section 3.2 in about twenty seconds on an ordinary workstation (64-bit Intel(R) Core(TM) i9-9900K CPU @ 3.60GHz with 16 cores).

A.2 Toy problem

This 1d example was introduced by Binois et al. [2018]. The mean function is $f(x) = 2 \times \exp(-30(x-0.25)^2 + \sin(\pi x^2)) - 2$ with a noise function of $r(x) = \frac{1}{3}\exp(\sin(2\pi x))$. Observations are simulated as $y \sim f(x) + \epsilon$, where $\epsilon \sim N(0, \sigma^2 = r(x))$. The training data comprise of 100 input locations equally spaced in $[0, 1]$, with 15 replicates observed upon each. Therefore, the total number of observations is $N = 15 \times 100 = 1500$. The simulated data can be represented as $D_N = (x_i, y_i), i = 1 \ldots N$. Let $\mathbf{X}_N$ and $\mathbf{Y_N}$ collect inputs and outputs, respectively. We then model D_N as $\mathcal{GP}(D_N)$ and use four fitted models:

1) a full hetereoscedastic GP fit from the `hetGP` R package [Binois and Gramacy, 2021], via `mleHetGP` as in Section 5.1;

2) a moments-based alternative using a full GP.

3) moments-based alternatives using our Vecchia method (Sections 3.2 and 3.3) with a) the predictive mean $\mu_n^{(v)}$ (Vecchia A) and b) the upper 95th quantile of the predictive mean $\mu_n^{(v)95}$ resulting from the second-moments fit, $\mathcal{GP}(D_n^{(v)})$, from Section 3.3 (Vecchia B).

The implementation of 2) was identical to Vecchia A, but the GPs fitted to first and second moments were full GP fits rather than Vecchia approximations, and models were fitted using the Gaussian covariance kernel. Joint inference for lengthscale and nugget parameters were conducted with `jmleGP`.

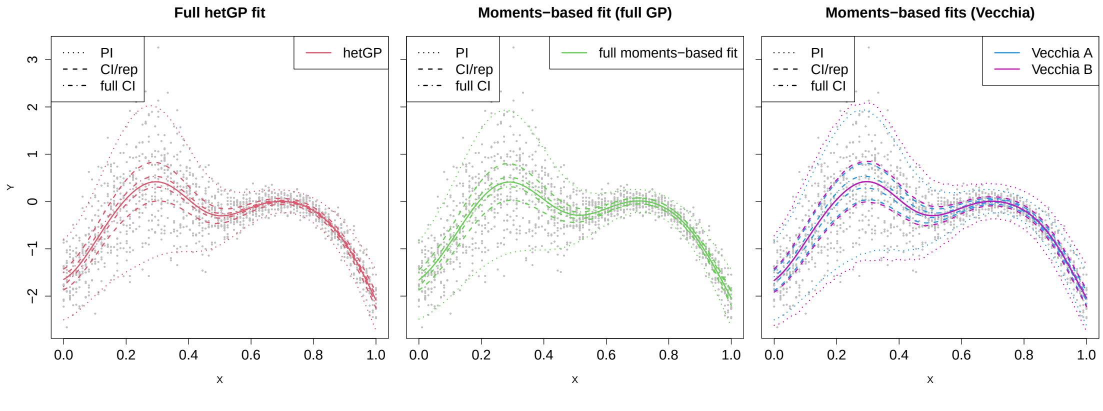

Figure 1: Fits for the four methods used in the toy example: hetGP (left), a moments-based alternative via a full GP fit (center), and moments-based alternatives using Vecchia (right). Vecchia A and B are as described above. 95% PI, CI, and “full CI” coverages are represented by dotted, dashed, and dot-dashed lines, respectively.

	PI	CI	CI-full
hetGP	0.94	0.97	1.00
Full moments-based fit	0.93	0.94	0.95
Vecchia A	0.94	0.96	0.98
Vecchia B	0.98	1.00	1.00

Table 1: 95% PIs and CIs for the four methods used in the toy example.

Simulated observations (gray dots), along with fits from all four models are shown in Figure 1, and 95% PIs and CIs are shown in Table 1. The predictive means from all four methods are nearly identical. We calculated CI coverage for methods 1)–3) above in two ways. For moments-based methods 2) and 3) we calculated standard errors for CI/PI calculations using $\hat{\sigma}_n^{\text{SK}}(\mathbf{x})^2$ from Eq. 11 in the main manuscript. For the hetGP implementation, our calculations were similar but involved dividing the diagonal of the predictive covariance matrix plus a varying nugget term by the number of replicates (15). We then determined the percentage of sample means that fell within CIs. To calculate PIs we did not divide by the number of replicates and report the percentage of data points that fell within the intervals. For this CI/PI calculation, performance is similar among the four methods with Vecchia B providing more conservative UQ by design. Vecchia B achieved 98% and 100% PI and CI coverage, respectively, while coverages of the other three methods ranged between 93 and 97% (Table 1, middle column).

However, this approach constitutes a CI for the average based on an iid assumption, which for this toy problem, is not true. When the spatial nature of the model is taken into account, the CIs become much narrower. When accounting for the spatial nature of the model in the toy problem, CIs should be compared to the *true mean*, not the sample average. For a real-world problem, we do not know the true mean, but because we know it in this example, it is possible to assess coverage in this way also. The standard error (SE) for these coverage calculations for methods 2) and 3) are given in Eq. (2):

$$\text{SE} = \sqrt{\frac{\mu_n^{(v)^*}(\mathbf{x})}{\text{dof}} + \frac{\sigma_n^{2(m)}(\mathbf{x})}{\text{dof}}}, \quad \text{where} \quad \text{dof} = \min_{i=1,\dots,n} \sum_{j=1}^{n} \Sigma(\bar{\mathbf{X}}_n)_{ij} \tag{2}$$

Here, $\Sigma_{ij} = k(q(\mathbf{x}_i, \mathbf{x}_j)) + g\mathbb{I}_{\{i=j\}})$, $k(\cdot)$ is the Gaussian covariance kernel, and $\Sigma(\bar{\mathbf{X}}_n)$ is $n \times n$, where $\bar{\mathbf{X}}_n$ is the matrix of unique input locations as defined in Section 3.2. In Eq. (2), the quantity $\mu_n^{(v)^*}(\mathbf{x})$ is derived from the *full* GP fit to $D_n^{(v)} = (\bar{\mathbf{X}}_n, \bar{\mathbf{S}}_n)$. The quantity $\sigma_n^{2(m)}(\mathbf{x})$ either represents the predictive variance of the full GP fit to $D_n^{(m)} = (\bar{\mathbf{X}}_n, \bar{\mathbf{Y}}_n)$ (the moments-based approach method 2) or predictive variance of the Vecchia-approximated fit to $D_n^{(m)}$ (moments-based approaches Vecchia A and B in method 3). While the SE in Eq. (2) is partially based on a full GP fit using a Gaussian covariance kernel, coverages for Vecchia A and B (which are fitted using a Matérn kernel with a smoothness parameter of 3.5) calculated this way still represent good estimates for the purposes of this example. Calculating true confidence intervals with the heteroscedastic GP implementation in method 1) requires only the diagonal of the "no nugget" predictive variance, which is returned by `mleHetGP`.

The final step involves a MC experiment in which CI coverages are calculated 100 times using Eq. (2), and the proportion of times the true mean falls within the CI is reported. These coverage values are given in the right-most column of Table 1. All four methods perform well in this regard, with all achieving at least 97% coverage.

A.3 Exercise with GLM runs

To benchmark our Vecchia-based workaround for SK we developed the following exercise to compare to true SK with respect to out of sample using a subset of NOAA-GLM runs. The goal was to have a large enough training exercise to stress our Vecchia calculations, while not being too large as to preclude expedient ordinary-GP calculations behind SK. We divided up one year's worth of NOAA-GLM hindcasts, including all depths and horizons, into five-day chunks. For each chunk we used a random $n_i = 16$ sample of NOAA-replicates for training, setting aside the other 15 for out-of-sample testing. In this way, each training data set was manageably sized at $n = 24,000$ (10 depths × 30 horizons × 5 days × 16 ensemble members) unique inputs. For each of the seventy-three spans of five days in 2022 we fit SK (Eqs. 9–10 in the main manuscript) and the Vecchia-approximated first- and second-moments, $\mu_n^{(m)}$ and $\sigma_n^{2(m)}$ for the mean and $\mu_n^{(v)}$ for full predictive uncertainty in Eq. (11), respectively. We saved out-of-sample RMSEs and coverages of 95% CIs and PIs. The calculation of CI coverage involved averages of the fifteen held-out replicates, at each unique input, whereas PIs used all of the hold-out replicates individually.

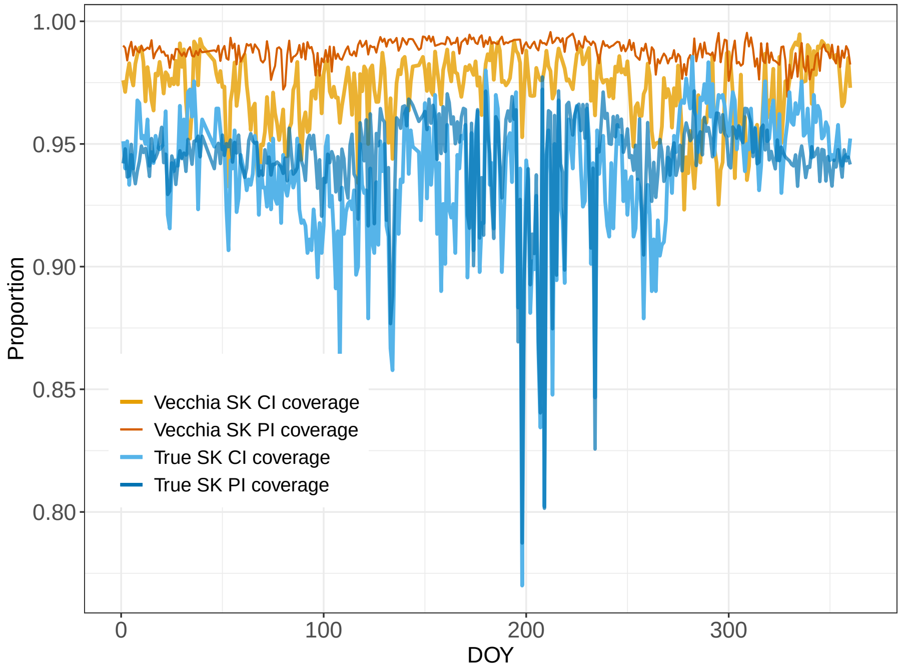

Figure 2: Vecchia approximation (Vecchia SK) and true SK 95% confidence and prediction interval (CI and PI, respectively) coverages over day of year (DOY) in 2022.

Figure 2 provides a compact summary of the results of that experiment. PI/CI coverages are shown as a proportion. PIs for both the approximate and true SK are fairly consistent throughout the year, whereas CIs fluctuate slightly more over DOY. Coverages of both true SK and approximate SK vary between DOY 200 and 250. We attribute the slightly higher variability in CI coverage of both methods to the relatively small 5-day training data size that was entertained. With only five days of examples to train on, there is not much diversity along the time axis in order to stabilize variances of first moments. However, average CI coverages over the entire year for both methods are close to the nominal value of 95%. True SK and approximate SK achieved 97.2 and 93.8% coverage, respectively, while average PI coverages for true SK and approximate SK are 94.5 and 99%. Overall, coverage of our Vecchia-based SK approximation (Eq. 11 in the main manuscript), via $\hat{\mu}_n^{\mathrm{SK}}(\mathbf{x})$ and $\hat{\sigma}^{\mathrm{SK}}(\mathbf{x})^2$ is slightly higher than that of SK, which is by design. Additionally, we found that the ratio of root mean-squared errors (RMSEs) from true SK and our Vecchia approximation, scaled by the range in temperatures at all depths and for all days in 2022, was near 0.035 for all days in 2022. This indicates that true SK and Vecchia are in agreement with respect to predictive accuracy.

B Iterative forecasting framework

Our forecasting is faithful to the brief description in Section 4.1, depicted diagrammatically in Figure 3, but is more accurately described by an operation that is performed all at once in a matter of minutes at midnight (00UTC), separating "yesterday" from "tomorrow". This is when NOAA forecasts for "tomorrow" (and the 29 subsequent days), and sensor measurements from "yesterday" become available. Yesterday's sensor measurements can be used to validate previous forecasts, and the metrics we prefer are discussed in Section 5.1 of the main manuscript. NOAA forecasts can be fed into GLM to produce 1–30 day-ahead lake temperature forecasts over each horizon and depth of interest. Those NOAA-GLM forecasts are then incorporated into the corpus of existing simulation training data, along with the sensor measurements. Model fitting for surrogates and bias correction can then again be carried out in an updating step, for example as described in Sections 3.3–3.4. Although there is potential for computational economy in priming aspects of new fits from previous ones, we do not generally bother because the Vecchia-based methods are fast. Updated surrogates may then be used to generate a 1–30 day-ahead bias-corrected forecast. The process repeats again at the next midnight as part of the iterative, near-term forecasting cycle [Dietze et al., 2018].

Forecasting accuracy and needs vary over the year during different meteorological and hydrological conditions and in response to changing manager needs. More frequent forecasts may be necessary in the fall to anticipate mixing events that are associated with water quality degradation [Thomas et al., 2020]. In comparison, when the lake water column is stratified in the summer and water temperatures are not rapidly changing, managers may require less frequent forecasts. Consequently, for our focal reservoir, it would be ideal if the updates to surrogates happened at a sub-daily time scale. However, in our effort to produce a holistic

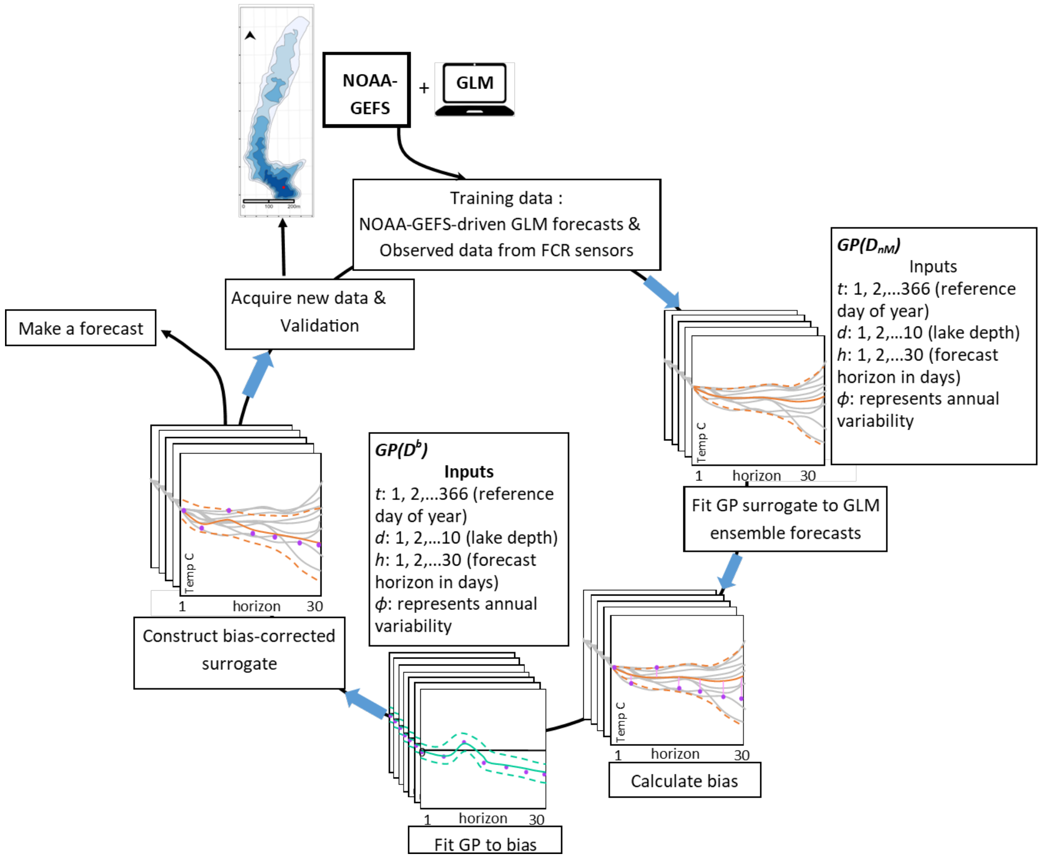
NOAA-GEFS
+
GLM
Training data :
NOAA-GEFS-driven GLM forecasts &
Observed data from FCR sensors
GP(D_nM)
Inputs
t: 1, 2,...366 (reference day of year)
d: 1, 2,...10 (lake depth)
h: 1, 2,...30 (forecast horizon in days)
ϕ: represents annual variability
Temp C
1 horizon 30
Fit GP surrogate to GLM ensemble forecasts
Calculate bias
Fit GP to bias
GP(D^b)
Inputs
t: 1, 2,...366 (reference day of year)
d: 1, 2,...10 (lake depth)
h: 1, 2,...30 (forecast horizon in days)
ϕ: represents annual variability
Construct bias-corrected surrogate
Make a forecast
Acquire new data & Validation

Figure 3: Iterative workflow for generating daily, 1–30 day-ahead forecasts of lake temperatures at 10 depths.

and robust forecasting apparatus, we opted for daily updates, which are generally sufficient for lake temperature forecasting.

C Scoring metrics

Given predictions $\hat{\mu} \equiv \hat{\mu}(\mathcal{X}) = (\hat{\mu}_1, \dots, \hat{\mu}_N)^\top$ and predictive variance $\Sigma = \Sigma(\mathcal{X})$ for out of sample $Y \equiv Y(\mathcal{X}) = (y_1, \dots, y_N)^\top$ score and RMSE may be calculated as follows:

$$\text{RMSE}(Y, \hat{\mu}) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{\mu}_i)^2}$$

$$\text{score}(Y, \hat{\mu}, \Sigma) = -\log|\Sigma| - (Y - \hat{\mu})^\top \Sigma^{-1}(Y - \hat{\mu}).$$

For score, Σ refers to the predictive variance of each method and is not fixed.

When validating for multiple horizons, and/or over multiple days, we aggregate accordingly. Score, RMSE, interval width and coverage are interesting because they provide a general sense of how the forecaster is performing. However, they are most useful when benchmarking against other, competing frameworks.

D Additional empirical results

Here we provide auxiliary empirical analysis for the forecasting experiment in Section 4.

D.1 Depth analysis

Figure 4 shows RMSE (A), score (B), coverage (C), and PI width for each depth (D), where results are aggregated over all horizons and days in the forecasting period. Similar to RMSE results over horizon, all competitors relying on GLM forecasts outperform OGP at deeper (>3m) depths, while OGP performs best at shallow depths (1–3m). NOAA-GLM forecasts can be highly variable at shallow depths, which may explain why forecasts from all GLM-surrogate assisted methods are less accurate than those of OGP at shallow depths. In contrast to RMSE results aggregated over horizon, GPBC (without ϕ) is consistently less accurate than GPBC (with ϕ). When examining RMSE results over depth, the improvement in accuracy provided by bias-correction in GPBC is not quite as dramatic as RMSE results summarized over horizon, but bias correction is still helpful.

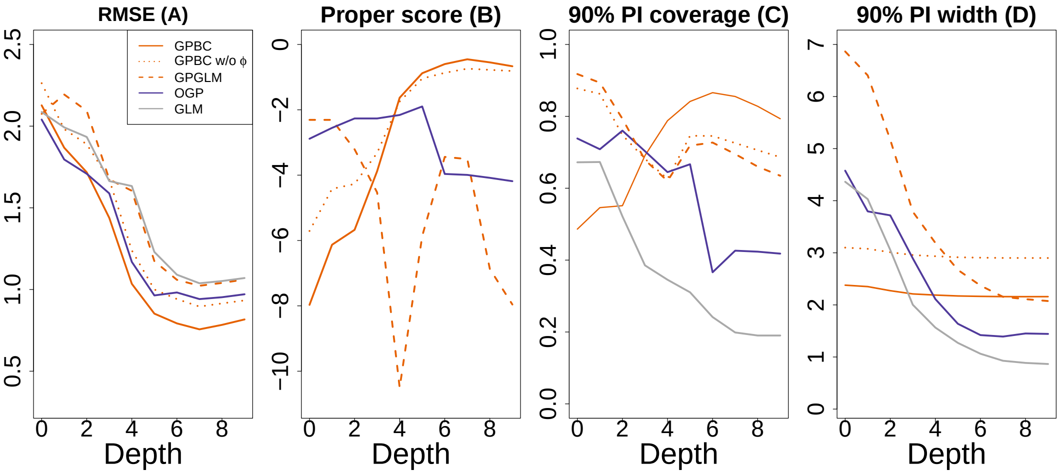

Figure 4: RMSE (°C), proper (log) score, and 90% PI coverage, and PI interval width from left to right respectively, summarizing from out-of-sample forecasts for all competing models over depth. Scores of GLM were omitted from the panel B because they were many orders of magnitude lower than other competitors.

The story is similar for proper scores. GPBC attains better (larger) scores than all other competitors at deeper depths. At shallow depths, GPBC (without ϕ) performs slightly better than its "with ϕ" counterpart. However at deeper depths, both variations of GPBC perform similarly. At surface and 1m depths, OGP and GPGLM perform best, but after 1m, GPGLM's scores fall steadily until 4m depth. As with results over horizon, raw GLM's scores fall beyond the y-limit range of the plot. Though raw GLM's accuracy over depth is similar to that of GPGLM, its UQ (discussed next) is poor, resulting in very low scores.

Coverage for all competitors except GPBC generally declines over depth, with some exceptions. While coverage of all other competitors decreases quickly after 1m depth, empirical coverage of GPBC actually increases steadily to 90% at 6m depth, declining to 80% at 9m depth. Even at 9m depth, GPBC still attains 80% coverage while other methods achieve only between 20 and 65% empirical coverage. GLM performs poorly overall, achieving a maximum empirical coverage of 68% only at surface depth. Finally, PI widths over depth for all competitors gradually decrease with increasing depth. A similar pattern exhibited in results over horizon is again apparent: PI widths of GPGLM are nearly three times as wide as those of GPBC at shallow depths. In addition, PI widths of GLM, and to some extent OGP, are much narrower than those of all other competitors at depths >3m. GLM's poor scores can be attributed to poor PI coverage over depth and inherent bias, and similarly, OGP's declining PI coverage over depth can likely be attributed to a decrease in PI width over depth.

D.2 Forecast accuracy over depth and horizon by day of year

Figure 5 focuses on forecasts over day of year (DOY) for two forecast horizons (1 and 30 days into the future). The top panel shows one-day ahead forecasts generated by GPGLM and GPBC, while the bottom panel shows forecasts at 30 days into the future at five depths. Observe in the top panels (one-day ahead) that GPGLM and GLM are generally in agreement with observations, lying nearly on top of one another. At thirty days out (bottom panels), the discrepancies between forecasts from GPGLM and GPBC are much more apparent, especially for deeper depths. At surface depths, the effect of bias-correction is not as obvious, but the improvement from bias-correction enhances accuracy at deeper depths between DOY 1 and 240.

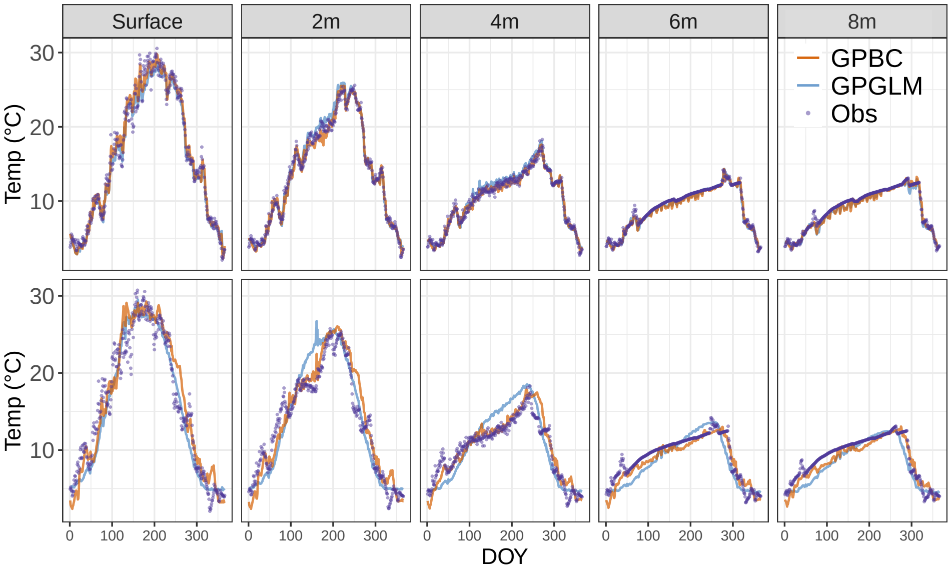

Figure 5: Forecasts from GPGLM (darker lines) and GPBC (ligher lines) generated during the forecasting exercise for horizon 1 (one day into the future, top) and horizon (30, 30 days into the future, bottom) over day of year (DOY) at surface, 2, 4, 6, and 8m depths (each panel denotes one depth). Observations are shown as dots.

D.3 Bias over day of year

Figure 6 shows how raw GLM bias (gray), surrogate GLM bias ("actual bias" via $\bar{\mathbf{Y}}_n^F - \hat{\mu}_n^{\mathrm{SK}}(\bar{\mathbf{X}}_n)$), and predicted bias vary over DOY in our out-of-sample forecasting exercise. Observe that while accuracy varies over horizon, the magnitude of bias from GLM depends even more strongly on the time of year a forecast is made [Figure 6]. Bias from GLM and its surrogate are similar. Predicted bias, which is used to correct it, follows a similar pattern, but it is not perfectly aligned with actual bias: it underestimates for the first half of the year, and then overestimates in the second half. Note that the predicted bias is not smooth, because bias for each day in the forecasting period was predicted iteratively, so each prediction was based upon a slightly different GP fit.

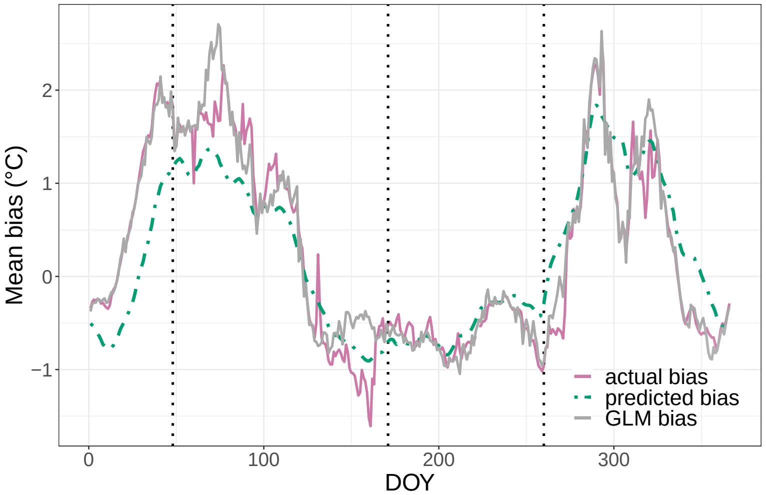

Figure 6: Bias (e.g., $\bar{\mathbf{Y}}_n^F - \hat{\mu}_n^{\text{SK}}(\bar{\mathbf{X}}_n)$ for out-of sample forecasts from the GLM surrogate (light, solid line), raw GLM (gray, solid line) and predicted bias (dot-dash line), averaged over depth and horizon for dates in the testing time period. Vertical black dashed lines denote example DOYs depicted in Figure 7.

D.4 Breaking out three days of forecasts

Figure 7 augments the visual in Figure 6 by breaking out horizon and depth contributions to bias for three particular days (indicated as vertical dashed lines in Figure 6). These three example days were chosen to display occasions when the bias-correction was correct in sign but not in magnitude (DOY 48 and 260) and generally correct in sign and magnitude (171). Consider DOY 48 first, in the left panels of Figure 7. Applying a bias correction (GPBC) improves the accuracy of forecasts compared to the GLM surrogate without a bias correction (GPGLM) with respect to observations (dots) at both 0 and 5m depths, although at 5m depth the forecast is still characterized by a cold bias. Moving to DOY 260 (right panel), the bias-correction actually worsens the forecast at the surface depth at after 15 days into the future. Finally, at day 171 (middle panel), forecasts at surface and 5m depths are generally improved by the bias-correction. These plots demonstrate that predicted bias applied to out-of-sample forecasts from GPBC mainly improves forecasts, although there are periods during the year when bias-correction over- or under-corrects or is of the wrong magnitude.

E Additional discussion

We suspect the main reason we lose to the climatological model (OGP) is a paucity of training data for estimating bias and the quality of NOAA forecasts used to drive GLM. We only have 2.5 years, meaning that we have seen fewer than three examples of yearly variation. Although OGP has access to the same amount of observational data, its simpler design – not

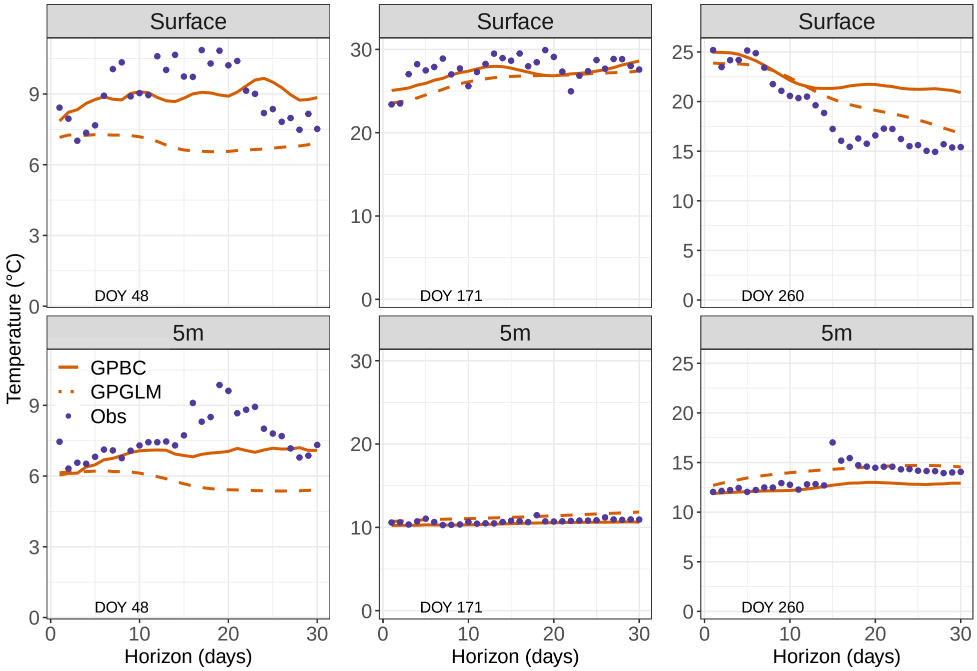

Figure 7: Out-of-sample predictive means of GPBC and GPGLM in solid and dashed lines, respectively, as well as observations as dark dots for 0m and 5m depth for the three days highlighted in Figure 6: 48, 171 and 260.

requiring the correction of poor model-based forecasts at longer horizons – helps to prevent over-fitting. GPBC can be modified so it is less weather dependent and behaves more like a climatological model by omitting the input ϕ (GPBC without ϕ). While it does not perform as well as the climatological model or GPBC with ϕ at shorter horizons, GPBC *with* ϕ is still our preferred model, because overall, it still performs better than GPBC without ϕ. Moreover, because the variability of GLM ensemble forecasts increases with forecast horizon, it is likely that the shape and degree of bias at longer horizons is more difficult to estimate compared to shorter horizons, where ensemble forecasts are more in agreement. Thus, bias correction at longer horizons may be less effective than at lower horizons without many more examples of yearly variation.

The quality of NOAA-GLM forecasts is largely dependent on the quality of NOAA forecasts themselves (assuming that GLM is well-calibrated). Weather forecast accuracy more than 10 days into the future declines quickly [Bauer et al., 2015]. Moreover, NOAA forecasted variables represent spatial averages over a 25km grid cell, much bigger than FCR for example, which may be too coarse to capture fine-scale variation in meteorological variables. Downscaling and bias-correction of NOAA forecasted variables could result in more accurate lake temperature forecasts from GLM, but additional post-processing of NOAA forecasts would take substantial computation time. Consequently, we decided that an investigation was beyond the scope of our current study. While forecasts of the climatological benchmark model were more accurate than GPBC at longer horizons, the UQ of those forecasts were generally poorer compared to those of GPBC.

It could be possible to create a hybrid forecasting model, in which the strength of $\phi(\lambda, t)$ is adjusted depending on the forecast horizon. Such an approach would combine the merits of GPBC with ϕ (better at shorter horizons) and GPBC without ϕ (more accurate at longer horizons). It is also worth noting that while OGP was only training on about 2.5 years of data, it performed exceedingly well at forecast horizons greater than 2 weeks. OGP shows that even very simple climatological models provide hard-to-beat benchmarks.

While the covariance for DOY should technically be periodic, we opted to use a standard (non-periodic) Gaussian covariance function for OGP. We investigated the use of a workaround to obtain a periodic fit, but we found that this did not result in any significant improvements to the non-periodic fit.

A way to extend this work would be to use GPBC as a likelihood for setting configuration parameters to GLM. In other words, to use the surrogate as a statistical calibration mechanism for the computer model, as opposed to relying on expert opinion using observed meteorological data, not NOAA forecasts.

We could also treat the derived input $\phi(t, \lambda)$, allowing us to distinguish forecasts among years, (i.e., the weather inputs) as a latent variable that could be estimated along with other quantities, thus allowing us to distinguish forecasts among years. Such an approach, which involves estimating a high-dimensional free quantity for all t and λ, is precluded, at the moment, by a paucity of sensor measurements. Perhaps once we have accumulated 10+ years of additional data , spanning more yearly environmental variability, the benefits of such an approach might outweigh the additional estimation risk it implies.

Another extension of this work could focus on extremes. Extreme events, such as heatwaves, rapid cooling events, or autumn mixing, are particularly critical for managing drinking water resources and mitigating water quality degradation. These events can lead to conditions such as hypoxia, algal blooms, or disruptions in water treatment processes. Accurately forecasting these extremes, as well as quantifying the uncertainty associated with them, is essential for operational decisions. For example, improved predictions during extreme mixing events could help managers adjust intake levels to minimize the impact on drinking water quality.

While our current methodology is designed for daily surrogate updates, incorporating extremes into the forecasting framework could involve future enhancements, such as prioritizing extremes in model training using weighted loss functions or cost-sensitive approaches, adopting adaptive surrogate models that increase temporal resolution (e.g., hourly) based on triggers such as rapid changes in meteorological conditions, exploring event-specific forecast evaluation metrics that better capture performance during rare but impactful scenarios.

References

P. Bauer, A. Thorpe, and G. Brunet. The quiet revolution of numerical weather prediction. *Nature*, 525(7567):47–55, 2015.

M. Binois and R. B. Gramacy. hetgp: Heteroskedastic Gaussian process modeling and sequential design in r. *Journal of Statistical Software*, 98:1–44, 2021.

M. Binois, R. B. Gramacy, and M. Ludkovski. Practical heteroscedastic Gaussian process modeling for large simulation experiments. *Journal of Computational and Graphical Statistics*, 27(4):808–821, 2018.

M. C. Dietze, A. Fox, L. M. Beck-Johnson, J. L. Betancourt, M. B. Hooten, C. S. Jarnevich, T. H. Keitt, M. A. Kenney, C. M. Laney, L. G. Larsen, et al. Iterative near-term ecological forecasting: Needs, opportunities, and challenges. *Proceedings of the National Academy of Sciences*, 115(7):1424–1432, 2018.

J. Guinness. Gaussian process learning via fisher scoring of vecchia's approximation. *Statistics and Computing*, 31(3):25, 2021.

A. Sauer, A. Cooper, and R. B. Gramacy. Vecchia-approximated deep Gaussian processes for computer experiments. *Journal of Computational and Graphical Statistics*, 32(3):824–837, 2023.

R. Q. Thomas, R. J. Figueiredo, V. Daneshmand, B. J. Bookout, L. K. Puckett, and C. C. Carey. A near-term iterative forecasting system successfully predicts reservoir hydrodynamics and partitions uncertainty in real time. *Water Resources Research*, 56(11): e2019WR026138, 2020.