
Technical Appendices and Supplementary Material for STACI: Spatio-Temporal Aleatoric Conformal Inference

A Theoretical Results

A.1 Spectral Covariance Approximation

Theorem 1 *The prior mean of the spatiotemporal covariance function of the discrete process in (1) equals the Matérn correlation with distance defined as in (2) for all J , and the point-wise prior variance decreases at rate J .*

Proof of Theorem 1: Given the frequencies ω_j and latent dimensions $L(\mathbf{s}, t)$, the covariance of

$$Z(\mathbf{s}, t) = \sum_{j=1}^J \cos[\omega_{s,j}^T \mathbf{s} + \omega_{t,j} t + \omega_{L,j}^T \mathbf{L}(\mathbf{s}, t)] a_j + \sin[\omega_{s,j}^T \mathbf{s} + \omega_{t,j} t + \omega_{L,j}^T \mathbf{L}(\mathbf{s}, t)] b_j, \quad (1)$$

averaging over the amplitudes (a_j and b_j) is (using the sum/difference identity)

$$\text{Cov}\{Z(s, t), Z(s', t')\} = \frac{\sigma^2}{J} \sum_{j=1}^J \cos[\omega_{s,j}^T (\mathbf{s} - \mathbf{s}') + \omega_{t,j} (t - t') + \omega_{L,j}^T \{\mathbf{L}(\mathbf{s}, t) - \mathbf{L}(\mathbf{s}', t')\}].$$

Define $\mathbf{h} = \{\mathbf{s} - \mathbf{s}', t - t', \mathbf{L}(\mathbf{s}, t) - \mathbf{L}(\mathbf{s}', t')\}$ and $M(\mathbf{h})$ as the Matérn correlation with distance

$$d^2 = \|\mathbf{s} - \mathbf{s}'\|^2 / \rho_s^2 + (t - t')^2 / \rho_t^2 + \sum_{l=1}^p [L(\mathbf{s}, t) - L(\mathbf{s}', t')]^2 / \rho_l^2. \quad (2)$$

Then treating $\omega_j \stackrel{iid}{\sim} MVT_\nu(0, D)$, the expected value of the covariance function is

$$\begin{aligned} \mathbb{E}[\text{Cov}\{Z(s, t), Z(s', t')\}] &= \frac{\sigma^2}{J} \sum_{j=1}^J \mathbb{E} \cos[\omega_{s,j}^T (\mathbf{s} - \mathbf{s}') + \omega_{t,j} (t - t') + \omega_{L,j}^T \{\mathbf{L}(\mathbf{s}, t) - \mathbf{L}(\mathbf{s}', t')\}] \\ &= \frac{\sigma^2}{J} \sum_{j=1}^J \mathbb{E} \cos(\omega_j \mathbf{h}) \\ &= \frac{\sigma^2}{J} J \int \cos(\omega \mathbf{h}) f(\omega) d\omega \\ &= \sigma^2 M(\mathbf{h}). \end{aligned}$$

Similarly, the second moment is

$$\begin{aligned}
\mathbb{E}[\text{Cov}\{Z(s, t), Z(s', t')\}^2] &= \mathbb{E}\left[\frac{\sigma^4}{J^2} \sum_{j=1}^J \sum_{k=1}^J \cos(\omega_j \mathbf{h}) \cos(\omega_k \mathbf{h})\right] \\
&= \frac{\sigma^4}{J^2} \mathbb{E}\left[\sum_{j=1}^J \cos(\omega_j \mathbf{h})^2 + \sum_{j \neq k}^J \cos(\omega_k \mathbf{h}) \cos(\omega_j \mathbf{h})\right] \\
&= \frac{\sigma^4}{J^2} \mathbb{E}\left[\sum_{j=1}^J \frac{1 + \cos(2\omega_j \mathbf{h})}{2} + \sum_{j \neq k}^J \cos(\omega_k \mathbf{h}) \cos(\omega_j \mathbf{h})\right] \\
&= \frac{\sigma^4}{J^2} \left[J + \frac{J}{2} M(2\mathbf{h}) + (J^2 - J) M(\mathbf{h})^2\right] \\
&= \frac{\sigma^4}{J} \left[1 + \frac{1}{2} M(2\mathbf{h}) + (J - 1) M(\mathbf{h})^2\right].
\end{aligned}$$

Thus the variance is

$$\begin{aligned}
\text{Var}[\text{Cov}\{Z(s, t), Z(s', t')\}] &= \mathbb{E}[\text{Cov}\{Z(s, t), Z(s', t')\}^2] - \mathbb{E}[\text{Cov}\{Z(s, t), Z(s', t')\}]^2 \\
&= \frac{\sigma^4}{J} \left[1 + \frac{1}{2} M(2\mathbf{h}) - M(\mathbf{h})^2\right].
\end{aligned}$$

Therefore, the approximation is centered on the Matérn covariance function with accuracy that increases with J .

B Additional Computational Details

B.1 Prior Settings

Here, we show the full hierarchical model for the deep learning approximation of the non-stationary spatio-temporal (ST) Gaussian Process (GP). Assume our observed data is $Y(\mathbf{s}, t)$ and our model is $Z(\mathbf{s}, t)$ with error $\epsilon(\mathbf{s}, t)$. Additionally, assuming INR architecture $\phi(\cdot)$ with weights \mathbf{W}_ϕ and GP hidden layer $\mathbf{x}_Z^{(h)}$, we have:

$$\begin{aligned}
\mathbf{L}(\mathbf{s}, t) &= \phi(\mathbf{s}, t) \\
\mathbf{x}_Z^{(h)}(\mathbf{s}, t) &= \left[\cos(\mathbf{W}_{Z,s}^{(h)} \mathbf{s} + \mathbf{W}_{Z,l}^{(h)} \mathbf{L}(\mathbf{s}, t) + \mathbf{W}_{Z,t}^{(h)} t), \sin(\mathbf{W}_{Z,s}^{(h)} \mathbf{s} + \mathbf{W}_{Z,l}^{(h)} \mathbf{L}(\mathbf{s}, t) + \mathbf{W}_{Z,t}^{(h)} t)\right] \\
Z(\mathbf{s}, t) &= \mathbf{W}_Z^{(h+1)} \mathbf{x}_Z^{(h)}(\mathbf{s}, t) \\
Y(\mathbf{s}, t) &= Z(\mathbf{s}, t) + \epsilon(\mathbf{s}, t) \\
\mathbf{W}_\phi | \alpha &\stackrel{iid}{\sim} N(0, \alpha I) \\
\mathbf{W}_{Z,s}^{(h)}, \mathbf{W}_{Z,l}^{(h)}, \mathbf{W}_{Z,t}^{(h)} | \nu, \rho_s, \rho_t, \rho_l &\stackrel{iid}{\sim} \text{MVT}(\nu, \rho_s, \rho_t, \rho_l) \\
\mathbf{W}_Z^{(h+1)} | \sigma^2 &\stackrel{iid}{\sim} N(0, \frac{\sigma^2}{J}) \\
\epsilon(\mathbf{s}, t) | \tau^2 &\stackrel{iid}{\sim} N(0, \tau^2),
\end{aligned} \tag{3}$$

For this study, we set $\alpha \sim \text{InvGamma}(1, 0.05)$, $\log(\nu) \sim N(0.5, 0.5)$, $\log(\rho_s) \sim N(-2, 1)$, $\log(\rho_t) \sim N(-1, 0.5)$, $\log(\rho_l) \sim N(-2, 1)$ and $\sigma^2, \tau^2 \sim \text{InvGamma}(0.1, 0.1)$. These give relatively uninformative priors for the hyperparameters of interest.

B.2 Model settings

We initialize each of the model architectures with the following:

- **STACI-FFNP**: INR layers: 5, INR layer width: 1024, INR activation function: GeLU, FFNP Frequency constant: 30, FFNP Frequency number: 1024, Latent dimensions: 128, GP layer width: 5000, Initialized models: 10
- **STACI-FFNG**: INR layers: 5, INR layer width: 1024, INR activation function: GeLU, FFNG σ : 5.0 (MSS), 1.0 (AOD), FFNG Encode size: 1024, Latent dimensions: 128, GP layer width: 5000, Initialized models: 10
- **STACI-ResMLP**: INR layers: 5, INR layer width: 1024, INR activation function: GeLU, Latent dimensions: 128, GP layer width: 5000, Initialized models: 10
- **Deep RF**: Hidden layers: 5, Layer width: 1024, Bottleneck width: 128, Spatial Kernel: Matérn, Temporal Kernel: Matérn, Initialized models: 10
- **Deep GP**: Hidden layers: 2, Layer width: 12, Kernel: Matérn($\frac{3}{2}$), Inducing points: 512
- **SVGP**: Inducing points: 3000, Kernel: Matérn($\frac{3}{2}$)

For the Conformal models, we chose between 30-80 nearest neighbors based on cross validation on a subset of the training observation. We use the AdamW optimizer with learning rate 1e-5 for STACI [1]. Deep GP and SVGP were implemented using the GPytorch package [2] and use the Adam optimizer [3] with learning rate 0.01. Deep RF was implemented using the provided code by [4] using Bayesian Optimization with parameter bounds updated to reflect the $[0, 1] \times [0, 1]$ spatial dimensions and learning rate 0.001. All models were trained for 15 epochs, aside from Deep RF which was trained for 1 epoch with 15 iterations.

B.3 Evaluation Metrics

- **Root Mean Square Error**: Given predictions $\hat{\mathbf{Y}} = \{\hat{Y}_1, \dots, \hat{Y}_n\}$ and observed values $\mathbf{Y} = \{Y_1, \dots, Y_n\}$, the RMSE is

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2}$$

- **Negative Log Likelihood**: Given predictions $\hat{\mathbf{Y}} = \{\hat{Y}_1, \dots, \hat{Y}_n\}$, estimated standard deviation $\hat{\sigma} = \{\hat{\sigma}_1, \dots, \hat{\sigma}_n\}$ and observed values $\mathbf{Y} = \{Y_1, \dots, Y_n\}$, the Negative Gaussian Log Likelihood, with π omitted, is calculated as

$$\text{NLL} = \frac{n}{2} \log(\hat{\sigma}) + \frac{1}{2\hat{\sigma}} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2.$$

- **Continuous Ranked Probability Score**: Given predictions $\hat{\mathbf{Y}} = \{\hat{Y}_1, \dots, \hat{Y}_n\}$, estimated standard deviation $\hat{\sigma} = \{\hat{\sigma}_1, \dots, \hat{\sigma}_n\}$ and observed values $\mathbf{Y} = \{Y_1, \dots, Y_n\}$, define $z_i = (Y_i - \hat{Y}_i)/\hat{\sigma}_i$, and let Φ and ϕ be the standard normal CDF and PDF. The CRPS averaged over n points is

$$\text{CRPS} = \frac{1}{n} \sum_{i=1}^n \left[\hat{\sigma}_i \left(z_i \{2\Phi(z_i) - 1\} + 2\phi(z_i) - \frac{1}{\sqrt{\pi}} \right) \right].$$

- **Interval score**: Given prediction intervals (L_i, U_i) , observed values $\mathbf{Y} = \{Y_1, \dots, Y_n\}$ and desired Type-1 error rate α , the interval score is

$$\text{IS} = \frac{1}{n} \sum_{i=1}^n \left\{ (U_i - L_i) + \frac{2}{\alpha} (L_i - Y_i)_+ + \frac{2}{\alpha} (Y_i - U_i)_+ \right\}.$$

The interval score balances interval width with coverage properties [5]. The first term penalizes intervals that are too wide while the second two terms penalize the coverage, raising the score if more observations lie outside the constructed interval.

B.4 Ablation on SVGD samples

Here, we perform an ablation study on the AOD dataset for number of SVGD samples on prediction metrics and computation time per epoch using the STACI-FFNP model. The number of samples, M , reflects the number of samples we use to approximate the posterior distribution for parameters. The results are shown in Table 1. We see that adding models keeps RMSE and NLL fairly consistent. However, the computational time increases significantly, taking 2.5 minutes per epoch.

Table 1: **Ablation on SVGD samples.** RMSE, NLL and computation time per epoch for $M = \{5, 10, 30\}$ samples.

Samples	RMSE	NLL	Time (s)
M = 5	0.559	0.041	31
M = 10	0.560	0.042	39
M = 30	0.561	0.044	155

References

- [1] Ilya Loshchilov, Frank Hutter, et al. Fixing weight decay regularization in adam. *arXiv preprint arXiv:1711.05101*, 5:5, 2017.
- [2] Jacob R Gardner, Geoff Pleiss, David Bindel, Kilian Q Weinberger, and Andrew Gordon Wilson. Gpytorch: Blackbox matrix-matrix gaussian process inference with gpu acceleration. In *Advances in Neural Information Processing Systems*, 2018.
- [3] Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- [4] Weibin Chen, Azhir Mahmood, Michel Tsamados, and So Takao. Deep random features for scalable interpolation of spatiotemporal data. In *The Thirteenth International Conference on Learning Representations*, 2024.
- [5] Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.