Bayesian Transfer Learning Approaches for Large-scale Spatiotemporal Problems

Luca Presicce

Ph.D. student in Statistics

University of Milano-Bicocca, Department of Economics, Management & Statistics

John Hopkins University, February 18, 2026

Who am I ? A short introduction

B.Sc. Statistical Sciences

University of Bologna, 2019

Structural models for the analysis of Italian employment rate

with Prof. Bianconcini Silvia (110/110 cum laude)

M.Sc. Statistical Sciences

University of Milano-Bicocca, 2021

Models for spatial network data with measurement error in covariates: a Bayesian approach

with INLA for road accidents

with Prof. Borgoni Riccardo (110/110 cum laude)

Ph.D. Statistics

University of Milano-Bicocca, 2025+

Bayesian Transfer Learning Approaches for Large-scale Spatiotemporal Problems

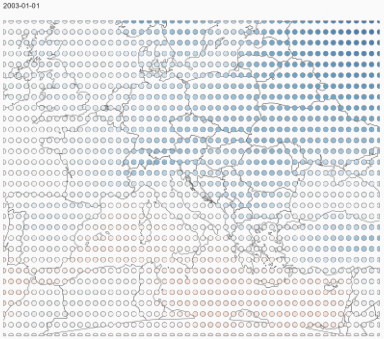
Prof.Banerjee Sudipto & Prof.Rigon Tommaso

Visiting researcher in 2023-2024 working with

Professor Sudipto Banerjee

University of California, Los Angeles (UCLA)

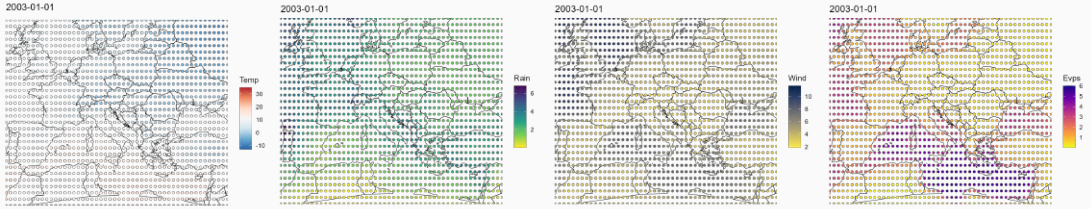
Why Should We Need Scalable Spatial and Spatiotemporal Models?



We may want to study weather components, as the surface temperature, collected over a hundreds of points

(small-scale problem - plenty of flexible modeling approaches)

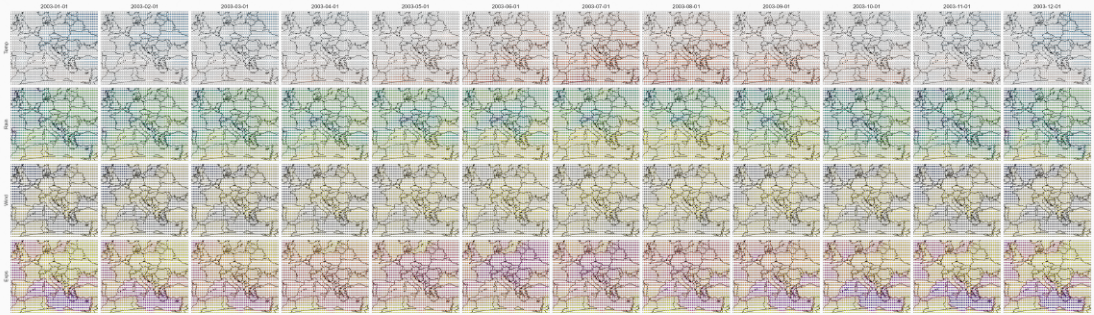
Why Should We Need Scalable Spatial and Spatiotemporal Models?



Meaningful weather components are multiple, leading thousands of points

(moderate-scale problem — standard modeling approaches may struggle)

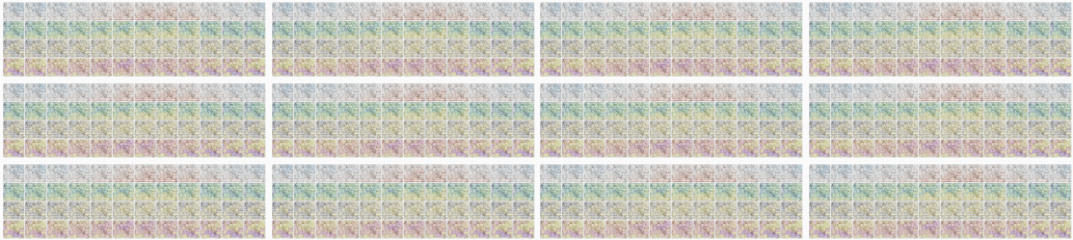
Why Should We Need Scalable Spatial and Spatiotemporal Models?



We may want to study spatiotemporal dependence of weather components, now we have tenth thousands of points

(above-average problem — possible modeling approaches reduce drastically)

Why Should We Need Scalable Spatial and Spatiotemporal Models?



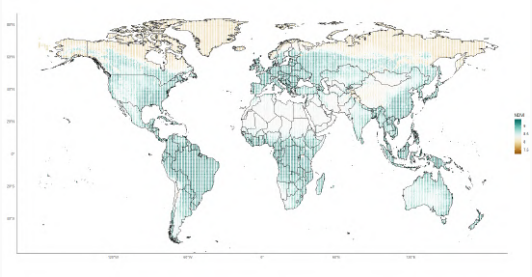
We can exploit multiple years of available multivariate data, with millions of points

(large-scale problem - Are there any feasible modeling approaches?)

Approaches for Large-scale Spatial Problems

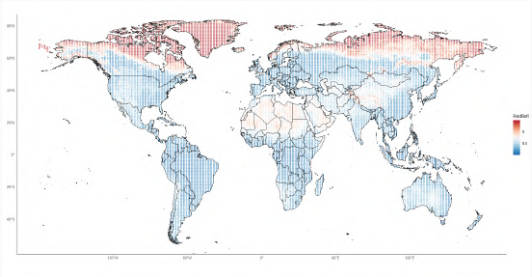
Research Goals

- Predict unobserved locations
- Leverage and study correlation



Data abundance from satellites

- Define suitable model
- Fit millions of record worldwide



Let's go full Bayes: using matrix-variate formulation for customary multivariate latent spatial regression

$$
\begin{aligned}
\mathbf{Y} \mid \boldsymbol{\beta}, \boldsymbol{\Omega}, \boldsymbol{\Sigma} &\sim \mathrm{MN}(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\Omega}, (\alpha^{-1} - 1)\mathbb{I}_n, \boldsymbol{\Sigma}) \\
\boldsymbol{\beta} \mid \boldsymbol{\Sigma} &\sim \mathrm{MN}(\mathbf{M}_0\mathbf{m}_0, \mathbf{M}_0, \boldsymbol{\Sigma}) \\
\boldsymbol{\Omega} \mid \boldsymbol{\Sigma} &\sim \mathrm{MN}(\mathbf{0}, \mathcal{R}(\mathcal{S}, \mathcal{S}; \phi), \boldsymbol{\Sigma}) \\
\boldsymbol{\Sigma} &\sim \mathrm{IW}(\boldsymbol{\Psi}_0, \nu_0) \\
\alpha &\sim p(\alpha) \\
\phi &\sim p(\phi).
\end{aligned}
$$

Where $\mathcal{S} = \{s_1, \ldots, s_n\}$ be a set of n locations yielding observations on q possibly correlated outcomes, collected into $(n \times q)$ matrix Y, while X is $(n \times p)$ matrix explanatory variables (full rank p).

Computational Limits

- Full MCMC: impractical (from $n \approx 10^3$)
- Variational/INLA: intensive (from $n \approx 10^4$)
- Memory/Time grows rapidly with n, q

Identifiability Problems

- Parameters α, ϕ weakly identifiable
- Strong prior assumptions required
- Strong human input required

Divide-&-Conquer

- Divide $\mathscr{D}$ into K subsets $\mathscr{D}_k$
- Work indipendently (in parallel) on each subset
- Combine local posteriors into a global one

Key Idea

Recover global posterior distribution assimilating K local posterior distributions

Model Averaging

- Define J model configurations $\mathscr{M}_j$
- Each $\mathscr{M}_j$ characterized as $\{\alpha_j, \phi_j\}$
- Combine conjugate posterior for all $\mathscr{M}_j$ $(j = 1, \ldots, J)$

Key Idea

Obtain local posterior distributions assimilating inference among J models

Local stacking (within each subset $\mathscr{D}_k$)

Within each subset $\mathscr{D}_k$, we compute the Bayesian predictive stacking weights $\{z_{k,j}\}$ as

$$\left(\hat{z}_{k,1}, \ldots, \hat{z}_{k,J}\right)^\top = \underset{z_k \in \mathscr{S}_1^J}{\arg\max} \frac{1}{n_k} \sum_{i=1}^{n_k} \log \sum_{j=1}^{J} z_{k,j}\, p\left(Y_{k,i} \mid \mathscr{D}_{k,[-l]}, \mathscr{M}_j\right),$$

- $Y_{k,i}$: i-th row of $Y_{k,[l]} \in \mathscr{D}_{k,[l]}$ (the l-th fold within the k-th dataset)
- $p\left(Y_{k,i} \mid \mathscr{D}_{k,[-l]}, \mathscr{M}_j\right)$: conditional posterior predictive
- n_k : number of location in $\mathscr{D}_k$
- L : number of folds for K-fold cross-validation
- $\mathscr{S}_1^J$: J-dimensional simplex.

Global stacking (between K subsets)

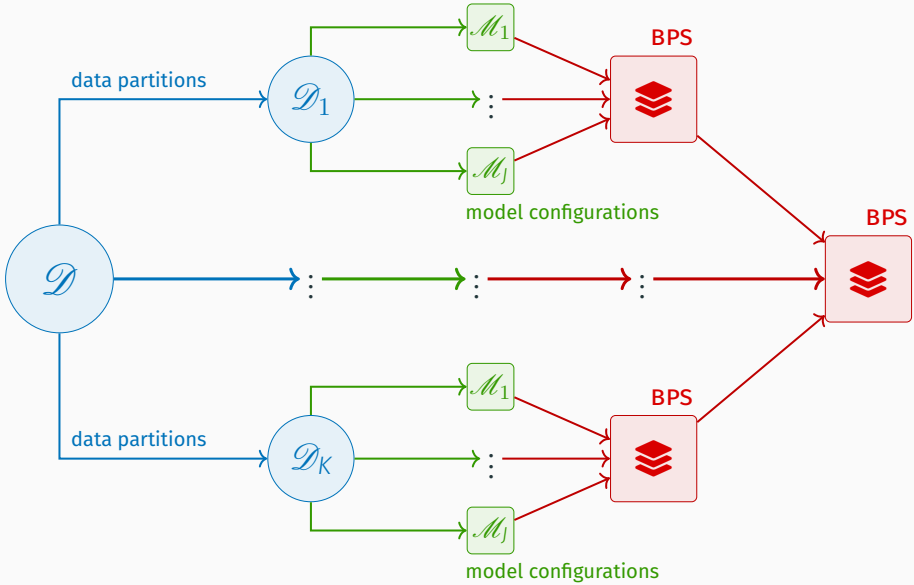
For inference on the full spatial dataset, we apply Bayesian predictive stacking again, which is equivalent to solving the following convex optimization problem:

$$\max_{w \in \mathcal{S}_1^K} \frac{1}{n} \sum_{i=1}^{n} \log \sum_{k=1}^{K} w_k \hat{p}\left(Y_i \; ; \; \mathscr{D}_{k,[-l]}\right) = \max_{w \in \mathcal{S}_1^K} \frac{1}{n} \sum_{i=1}^{n} \log \sum_{k=1}^{K} w_k \sum_{j=1}^{J} \hat{z}_{k,j} p\left(Y_{k,i} \mid \mathscr{D}_{k,[-l]}, \mathscr{M}_j\right).$$

Once get Bayesian predictive stacking weights $\hat{w} = \left\{\hat{w}_k\right\}_{k=1}^{K}$, estimation of any posterior distribution of interest is achieved as

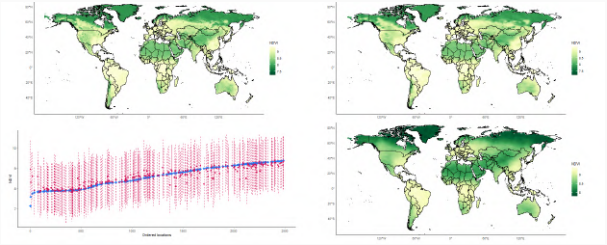
$$\hat{p}(\cdot \; ; \; \mathscr{D}) = \sum_{k=1}^{K} \hat{w}_k \sum_{j=1}^{J} \hat{z}_{k,j} \; p\left(\cdot \mid \mathscr{D}_k, \mathscr{M}_j\right) \; .$$

Double Stacking approach: conceptualization



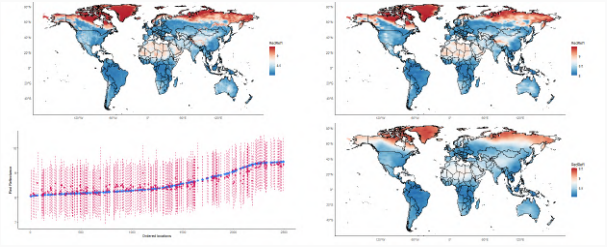
Run Time

- $n = 1,000,000$ multivariate observations
- ≈ 20 minutes on standard laptop



Predictive assessment

- Lower RMSPE than (feasible) competitor
- High empirical predictive coverage $> 95\%$



Bayesian Models

- Not possible to benchmark multivariate Bayesian spatial models
- Only feasible alternative: non-spatial multivariate Bayesian model
- We halved RMPSE, similar estimates for correlation

AI/ML Models

- Benchmarks: **GBM, Deep NNs, AutoML**
- We halved RMSPE, comparable running time
- Standard AI pipelines are univariate
- No uncertainty quantification

Key Advantages

- Fast Bayesian inference and Uncertainty quantification
- Automated with minimal human input
- Excellent predictive performance
- Scalable to large datasets with limited resources

Key Point

Fast, and reliable predictions and uncertainty quantification for multivariate spatial random fields with minimal input.

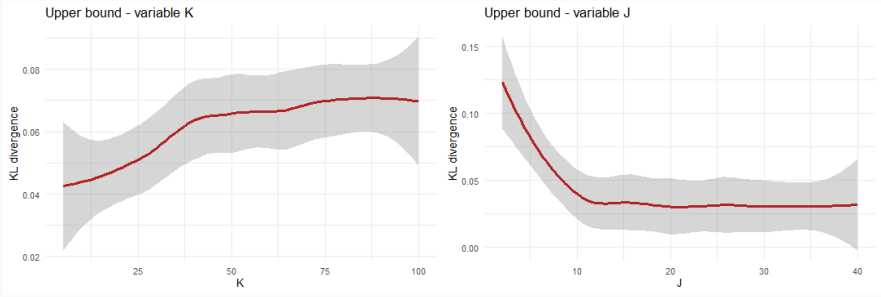
Theoretical Guarantees: Upper bound for KL divergence between DBPS and true predictive

We focus on the reversed Kullback-Leibler divergence between the DOUBLE BPS posterior predictive and the true one, deriving the following upper bound

$$D_{KL}\left(\hat{P} \,\|\, P_t\right) \leq \log \prod_{k=1}^{K} \left\{ \sum_{k=1}^{K} \hat{w}_k \sum_{j=1}^{J} \hat{z}_{k,j}\, \mathbb{E}_{\hat{p}_{k,j}} \left[\frac{\sum_{j=1}^{J} \hat{p}(y \mid \mathscr{D}_k, \mathscr{M}_j)}{p_t(y \mid \mathscr{D})} \right] \right\}^{\hat{w}_k}$$

where $\hat{P}$ and P_t denote probability measures with probability density $\hat{p}(\,\cdot\,; \mathscr{D})$, and $p_t(\,\cdot \mid \mathscr{D})$, and $\hat{p}_{k,j} = \hat{p}(\,\cdot \mid \mathscr{D}_k, \mathscr{M}_j)$.

Monte Carlo approximation allows study $D_{KL}\left(\hat{P} \,\|\, P_t\right)$ varying the number of subsets (K) and the number of candidate models (J)



Key Limitations

- Not full Bayes - no posterior inference on α, ϕ
- Prediction oversmoothing
- Results depend on partitioning method
- Results depend on single partition

Key Point

Trade-off between scalability and model richness: exploit full conjugacy at cost to lose full Bayes inference on spatial parameters

Approaches for Large-scale Spatiotemporal Problems

New Challenge: Time

- Data arrives sequentially over time
- Need for forecasting and online learning
- Temporal dependence structures
- High-dimensional in both space and time

Key Point

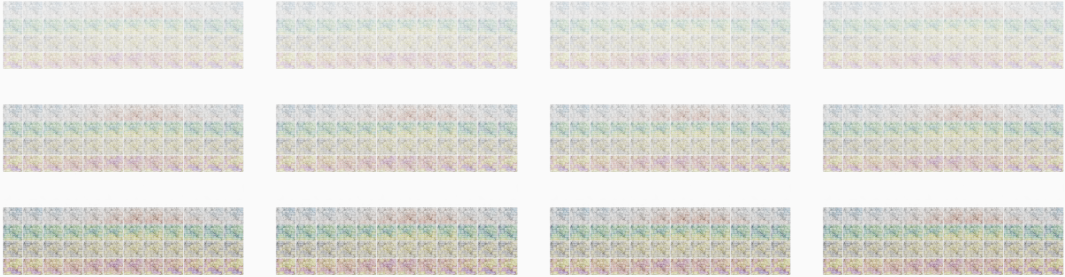
Develop Bayesian transfer learning approach to handle both space and time-varying stochastic processes while maintaining scalability

Research Goals

- Forecast and interpolate climate determinants
- Leverage and study correlation along time

Copernicus Climate Atlas Data

- Define multivariate spatiotemporal model
- Fit thousands of spacetime records



Let $\mathbf{Y} = \{\mathbf{Y}_t : t \in \mathcal{T} \subset \mathbb{N}\}$ be a spatiotemporal tensor of outcomes: $\mathbf{Y}_t$ $(n \times q)$ matrix with n fixed locations $\mathcal{S} = \{s_1, \ldots, s_n\}$, for q correlated outcomes.

A matrix-variate dynamic linear model can be represented as:

$$\mathbf{Y}_t = \mathbf{F}_t\boldsymbol{\Theta}_t + \boldsymbol{\Upsilon}_t, \qquad \boldsymbol{\Upsilon}_t \sim \mathrm{MN}(\mathbf{0}, \mathbf{V}_t, \boldsymbol{\Sigma})$$

$$\boldsymbol{\Theta}_t = \mathbf{G}_t\boldsymbol{\Theta}_{t-1} + \boldsymbol{\Xi}_t, \qquad \boldsymbol{\Xi}_t \sim \mathrm{MN}(\mathbf{0}, \mathbf{W}_t, \boldsymbol{\Sigma})$$

Observation Equation:

- $\mathbf{Y}_t$: Observed matrix at time t
- $\mathbf{F}_t$: Design matrix
- $\boldsymbol{\Theta}_t$: State matrix (latent)

Latent State Equation:

- $\boldsymbol{\Theta}_t$: State matrix at time t
- $\mathbf{G}_t$: Transition matrix
- $\boldsymbol{\Theta}_{t-1}$: State matrix at time $t-1$

We cast matrix-variate DLM as multivariate autoregressive latent spatial regression

$$\begin{aligned}
\mathbf{Y}_t \mid \mathbf{B}_t, \boldsymbol{\Omega}_t, \boldsymbol{\Sigma} &\sim \mathrm{MN}(\mathbf{X}_t\mathbf{B}_t + \boldsymbol{\Omega}_t, \mathbf{V}_t(\alpha), \boldsymbol{\Sigma}) \\
\mathbf{B}_t \mid \boldsymbol{\Sigma} &\sim \mathrm{MN}(\mathbf{B}_{t-1}, \mathbf{W}_t^{(B)}, \boldsymbol{\Sigma}) \\
\boldsymbol{\Omega}_t \mid \boldsymbol{\Sigma} &\sim \mathrm{MN}(\boldsymbol{\Omega}_{t-1}, \mathcal{R}_t(\mathcal{S}, \mathcal{S}; \phi), \boldsymbol{\Sigma}) \\
\boldsymbol{\Sigma} &\sim \mathrm{IW}(\boldsymbol{\Psi}_t, \nu_t) \\
\alpha &\sim p(\alpha) \\
\phi &\sim p(\phi).
\end{aligned}$$

Spatial Structure:

- $\mathbf{W}_t = \begin{bmatrix} \mathbf{w}_t^{(B)} & \mathbf{0} \\ \mathbf{0} & \mathcal{R}_t(\mathcal{S}, \mathcal{S}; \phi) \end{bmatrix}$
- $\mathcal{R}_t(\mathcal{S}, \mathcal{S}; \phi)$: Spatial correlation
- $\mathbf{V}_t(\alpha) = (\frac{1-\alpha}{\alpha})\mathbb{I}_n$: Nugget discontinuity
- $\alpha \in (0, 1)$: proportion of spatial variability

Dynamic Components:

- $\boldsymbol{\Theta}_t = [\mathbf{B}_t^\top : \boldsymbol{\Omega}_t^\top]^\top$
- $\mathbf{B}_t$: regression coefficients
- $\boldsymbol{\Omega}_t$: Latent spatial process

Computational Limits

- Full MCMC: impractical for moderate n, T
- SPDE: numerically intensive
- Memory/Time grows rapidly with n, q, T

Identifiability Problems

- Parameters α, ϕ weakly identifiable
- Strong prior assumptions required
- Strong human input required

Bayesian Predictive Stacking

- Define J model configurations $\mathscr{M}_j$
- Each $\mathscr{M}_j$ characterized as $\{\alpha_j, \phi_j\}$
- Combine conjugate posterior for all $\mathscr{M}_j$ ($j = 1, \ldots, J$)

Key Point

Avoid iterative or simulation-based approaches averaging out α, ϕ over J couples

Advantages

- Exploit conjugate inference
- Efficient computation and sampling

Disadvantages

- Breaks posterior-to-prior updates
- Overlooks temporal dynamics

⚠ We devise two workaround controlling posterior-to-prior propagation!

Variational Propagation

- Compute stacking weights $\hat{w}_t$ using BPS
- Obtain stacked posterior inference using
  $\hat{p}(\cdot\,; \mathscr{D}_{1:t}) = \sum_{j=1}^{J} \hat{w}_{j,t}\, p(\cdot \mid \mathscr{D}_{1:t}, \mathscr{M}_j)$
- Approximate $\hat{p}(\cdot\,; \mathscr{D}_{1:t})$ for conjugate update

Key Point

Use a variational approximation of stacked posterior to posterior-to-prior conjugate family

We approximate the mixture stacked posterior with a tractable conjugate distribution, minimizing Kullback-Leibler divergence:

$$D_{KL}\left(\sum_{j=1}^{J} \hat{w}_{t,j}\, p(\boldsymbol{\Theta}_t, \boldsymbol{\Sigma} \mid \mathscr{D}_{1:t}, \mathscr{M}_j) \,\middle\|\, \hat{p}_{KL}(\boldsymbol{\Theta}_t, \boldsymbol{\Sigma}\,;\, \mathscr{D}_{1:t})\right)$$

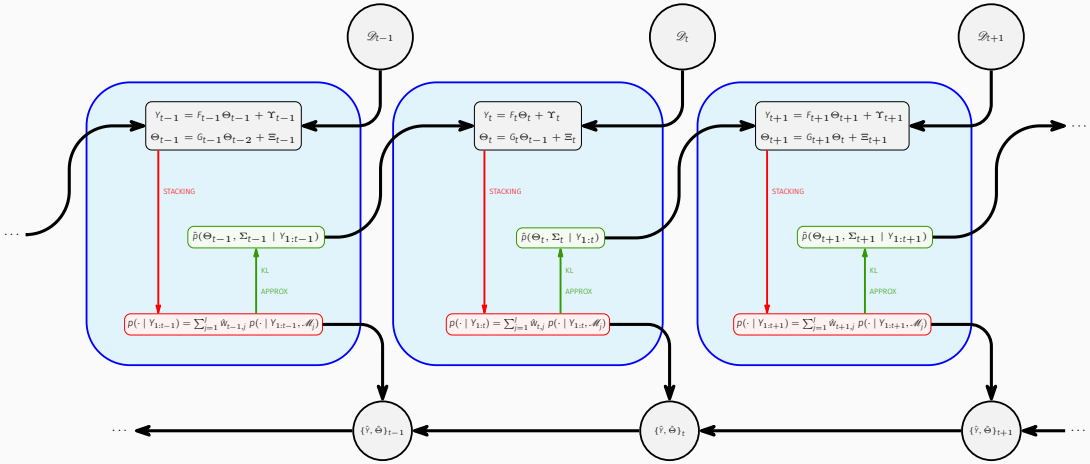
Obtain the variational approximating posterior distribution analytically

$$\hat{p}_{KL}(\Theta_t, \Sigma\,;\, \mathscr{D}_{1:t}) = \mathrm{MNIW}(\Theta_t, \Sigma \mid \tilde{m}_t, \tilde{C}_t, \tilde{\Psi}_t, \tilde{\nu}_t) \tag{1}$$

With minimizing parameters:

- $\tilde{m}_t = \sum_{j=1}^{J} \hat{w}_{t,j} m_t^{(j)}$
- $\tilde{C}_t = \sum_{j=1}^{J} \hat{w}_{t,j}\left[C_t^{(j)} + (m_t^{(j)} - \tilde{m}_t)^\top \Sigma^{-1} (m_t^{(j)} - \tilde{m}_t)\right]$
- $\tilde{\Psi}_t = \tilde{\nu}_t \left[\sum_{j=1}^{J} \hat{w}_{t,j} \nu_t^{(j)} \Psi_t^{-1(j)}\right]^{-1}$
- $\tilde{\nu}_t = \sum_{j=1}^{J} \hat{w}_{t,j} \nu_t^{(j)}$

Variational Propagation: Conceptualization



Parallel Propagation

- Work in parallel for J conjugate flows
- Compute stacking weights $\hat{w}_t$ using DYNBPS
- Obtain stacked posterior inference using
  $\hat{p}(\cdot\,; \mathscr{D}_{1:t}) = \sum_{j=1}^{J} \hat{w}_{j,t}\, p(\cdot \mid \mathscr{D}_{1:t}, \mathscr{M}_j)$

Key Point

Fit parallel conjugate models, then assimilate inference using DYNBPS

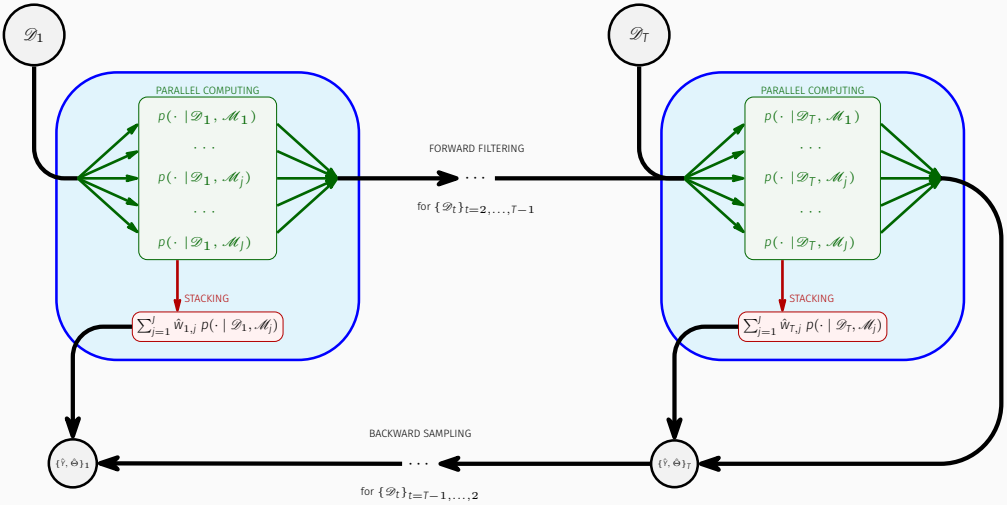
Dynamic BPS (across time instants)

At any time instant $\tau \in \mathcal{T} \subset \mathbb{N}$, we compute dynamic Bayesian predictive stacking weights $\{\hat{w}_i^{(\tau)}\}$ as

$$\big(\hat{w}_{i,1}^{(\tau)}, \ldots, \hat{w}_{i,J}^{(\tau)}\big)^\top = \underset{w_i^{(\tau)} \in \mathscr{S}_1^J}{\operatorname{argmax}} \frac{1}{(\tau-1)} \sum_{t=1}^{\tau-1} \log \sum_{j=1}^{J} w_{i,j}^{(\tau)}\, p(Y_{t+1,i} \mid \mathscr{D}_{1:t}, \mathscr{M}_j),$$

- $Y_{t,i}$: i-th row of $Y_t \in \mathscr{D}_{1:t}$ (available information up to t)
- $p(Y_{t,i} \mid \mathscr{D}_{t-1}, \mathscr{M}_j)$: conditional one-step ahead posterior predictive
- $(\tau - 1)$: training time-points for leave-future-out cross-validation
- $\mathscr{S}_1^J$: J-dimensional simplex

Parallel Propagation: Conceptualization



Variational

- Restore conjugate posterior-to-prior updates
- Over-inflates variance in $\mathscr{M}$-open settings
- KL approximation only acts as prior distirbution
- Static Bayesian predictive stacking weights

Parallel

- Fit conjugate models in parallel
- Perform better in $\mathscr{M}$-open settings
- Computationally more efficient
- Dynamic Bayesian predictive stacking weights

Data Description

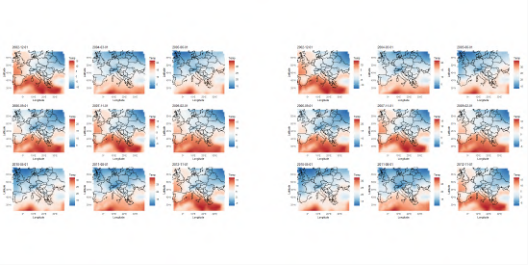
- Space: $1,500$ locations (central Europe)
- Time: 144 months (from Dec 2002 to Dec 2014)
- Outcomes: temperature, precipitation, wind, evaporation
- Covariates: cloud-coverage, solar radiation, sea-level pressure, near-surface humidity, 11 monthly-dummies

Prediction Tasks

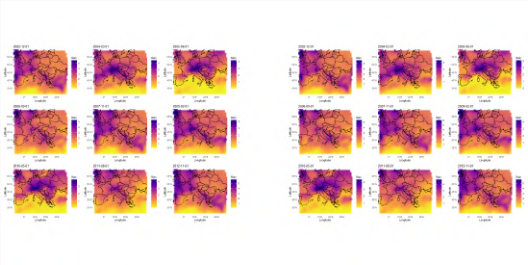
We analyze $1,500 \times 144 \times 4 = 864,000$ spacetime points using parallel propagation and DYNBPS

Results: One-Step-Ahead Forecasting

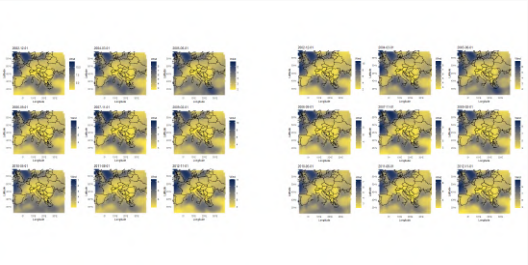
Temperature



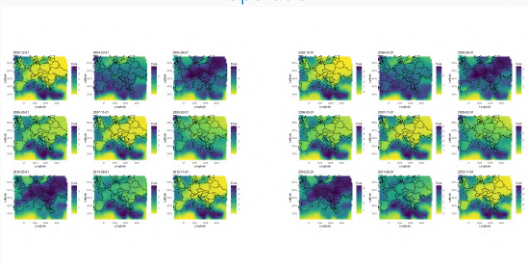
Precipitation



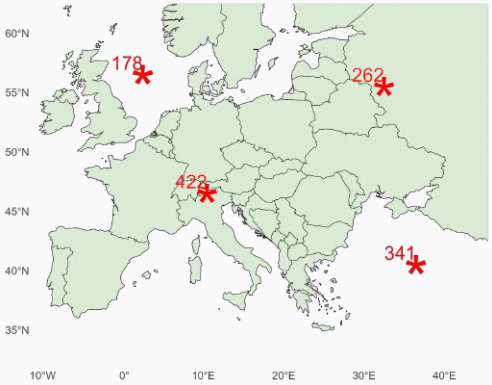
Wind

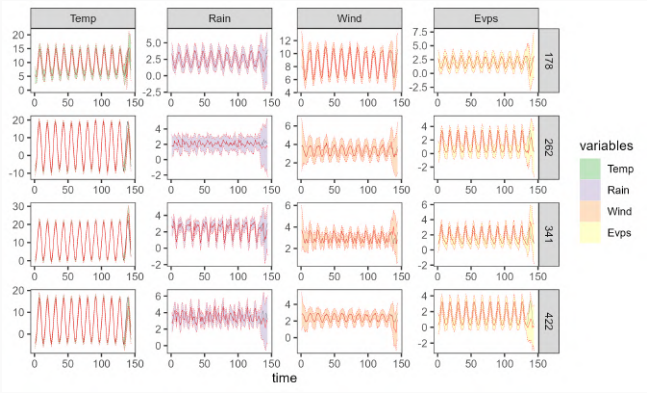


Evaporation



Results: One-Step-Ahead Forecasting

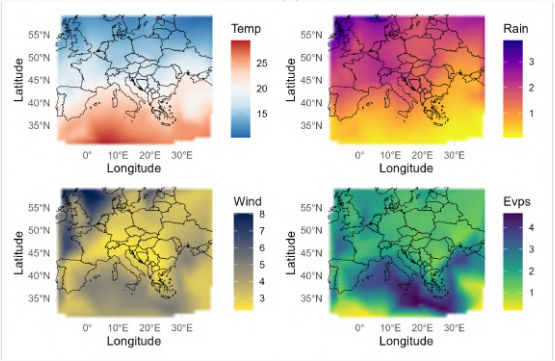




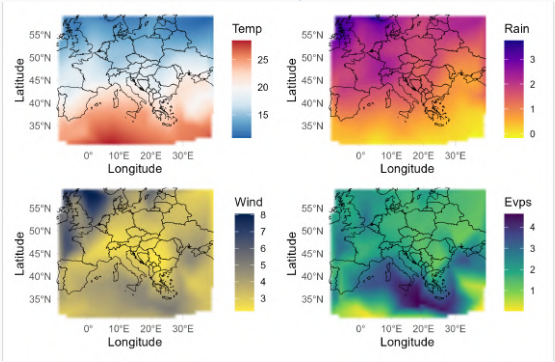
One-step ahead monthly forecast for randomly selected locations

Results: Spatial Interpolation

True



DYNBPS



Spatial surface interpolation at unobserved time (out of sample)

Remarks on Bayesian Transfer Learning Approaches for Large-scale Problems

Methodological

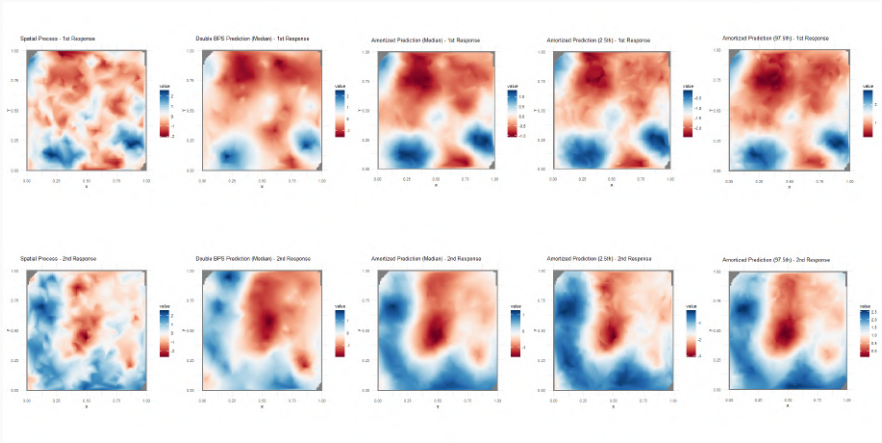
- Double Bayesian predictive stacking (DBPS)
- KL upperound for DBPS predictive
- Dynamic extension of stacking (DYNBPS)
- Two spacetime propagation strategies

Practical

- Scalable to multivariate massive datasets
- Accessible with limited resources
- Automated with minimal input
- Flexible open-source software

Bayesian Transfer Learning

- Use DBPS as generator of synthetic data
- Enables scaling amortized inference in new domains
- Provide instantaneous uncertainty quantification



Limitations

- Separable covariance assumption
- No full Bayes approaches
- Arbitrary extension of parameters grid
- Dependence on partition schemes

Future Work

- Non-separable covariance structures
- Exponential families extension
- Amortized inference formalization
- Dependent stacking weights

Still much work needs to bridge the research gap!

Strengths

- Scalable and largely automated
- Predictive tool for massive, multivariate data
- Flexible frameworks with possible extensions

Considerations

- Trade-off between scalability and richness
- Oversmoothing must be taken into account
- Separability limits covariance flexibility

> Far better an approximate answer to the right question, than an exact answer to the wrong question - J. Tukey

R Packages

👉 spBPS: Bayesian Predictive Stacking for Scalable Geospatial Transfer Learning

👉 spFFBS: Spatiotemporal Propagation for Multivariate Bayesian Dynamic Learning

Thanks for attention!

See more of my work and publications here:

https://lucapresicce.github.io

Scan the QR code or visit the link to learn more.