A Bayesian Factor Analysis for Temperature Projections

Shaleni Kovach and Richard L. Smith

EnviBayes workshop, Texas A&M

October 30, 2025

THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL

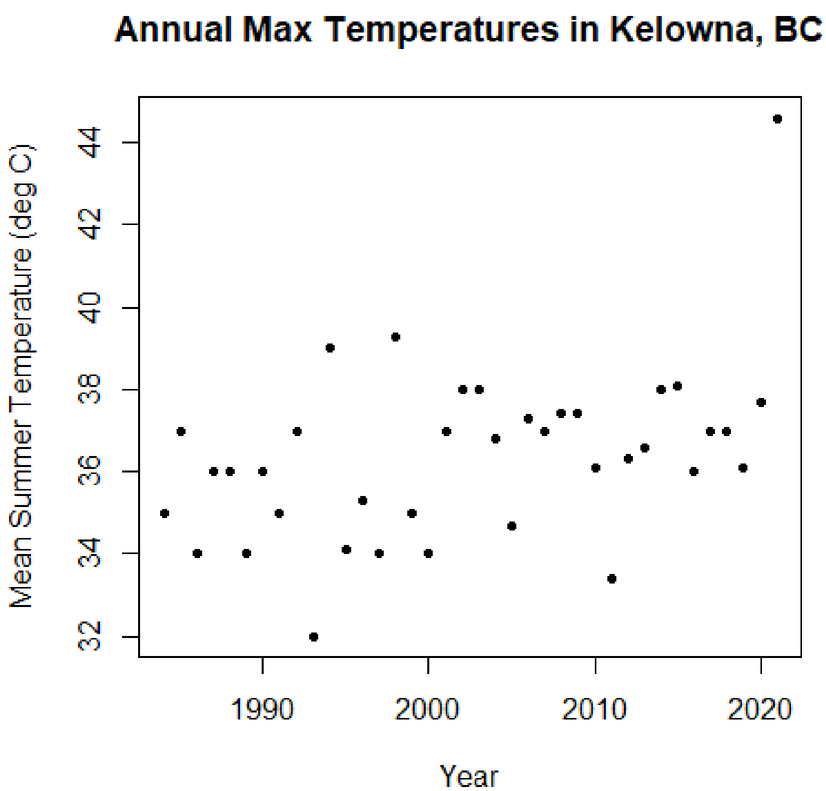
1

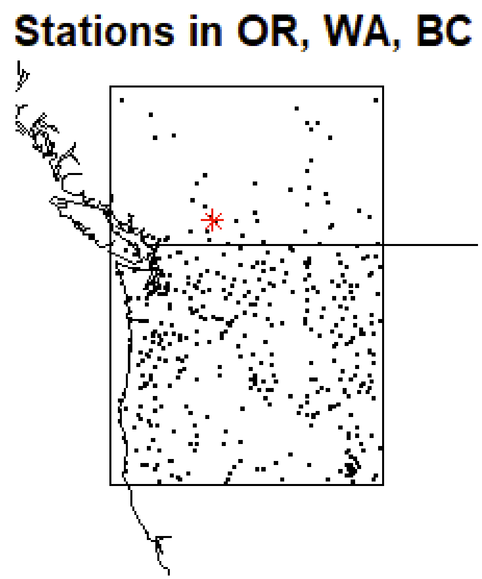
A Paradigm for Statistical Climate Research

- Define an event of interest
- Build a statistical model to relate that event to a "climate" variable*
- Calculate the same climate variable in a climate model (preferably, use several climate models)
- Project forwards using any of the "Shared Socioeconomic Pathways" (SSPs)
- Summarize the results as projected probabilities or magnitudes for the event of interest, with uncertainty bounds

*For the purpose of this talk, a climate variable is something that can be calculated from a climate model, such as mean annual or summer temperature over a specified region of the earth

2021 Heatwave in Pacific Northwest





Preliminary Analysis of Kelowna Data

- Objective: fit Generalized Extreme Value (GEV) model to data from 1984–2020, use it to calculate extreme events probabilities in 2021 and beyond

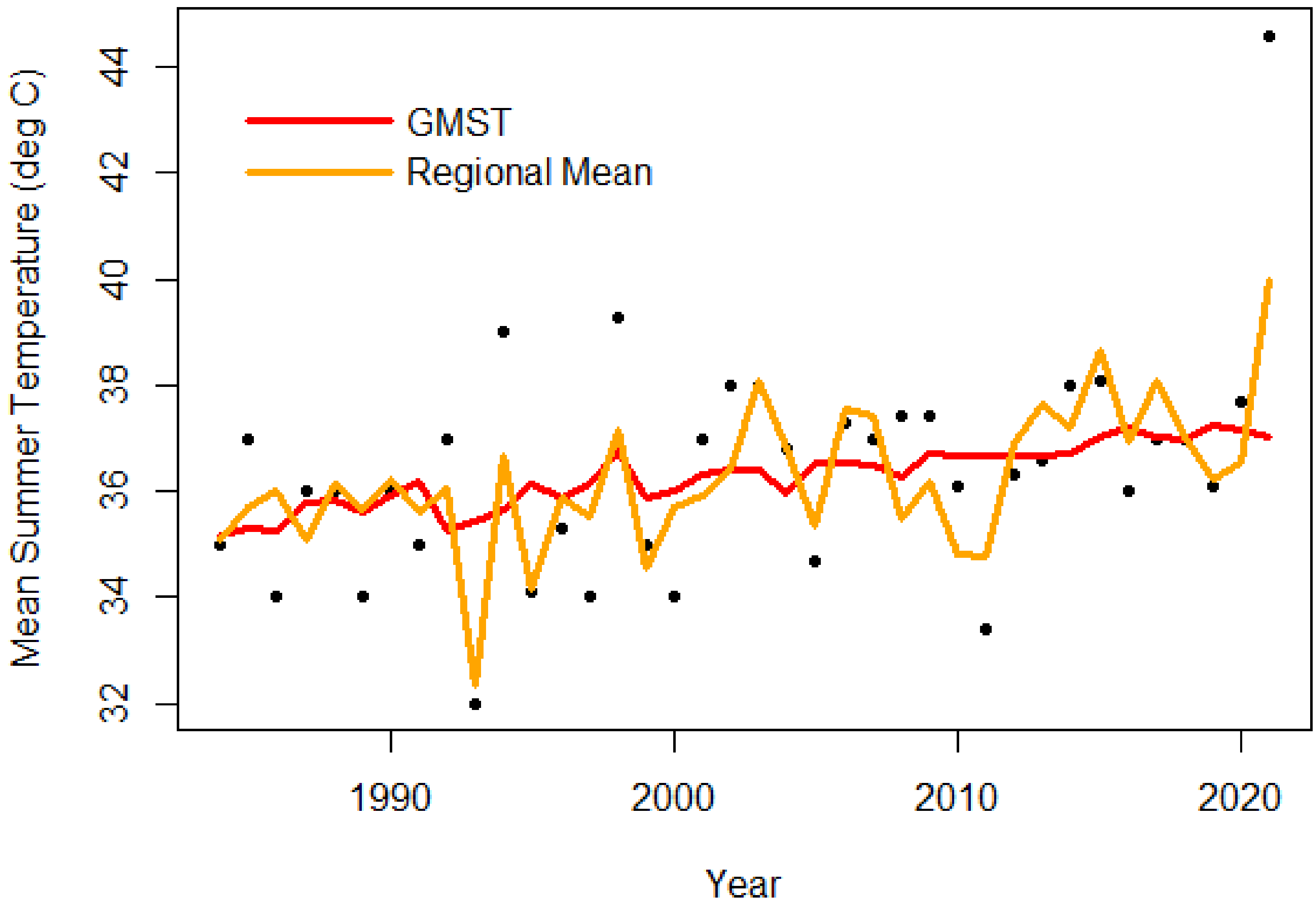
- Use a covariate: either GMST or a regional summer temperature mean (40–55 ^{o}N, 110–125 ^{o}W)

- NLLH=69.7 (no covariate) or 67.5 (GMST) or 55.7 (regional mean) — clear preference for regional mean as a covariate

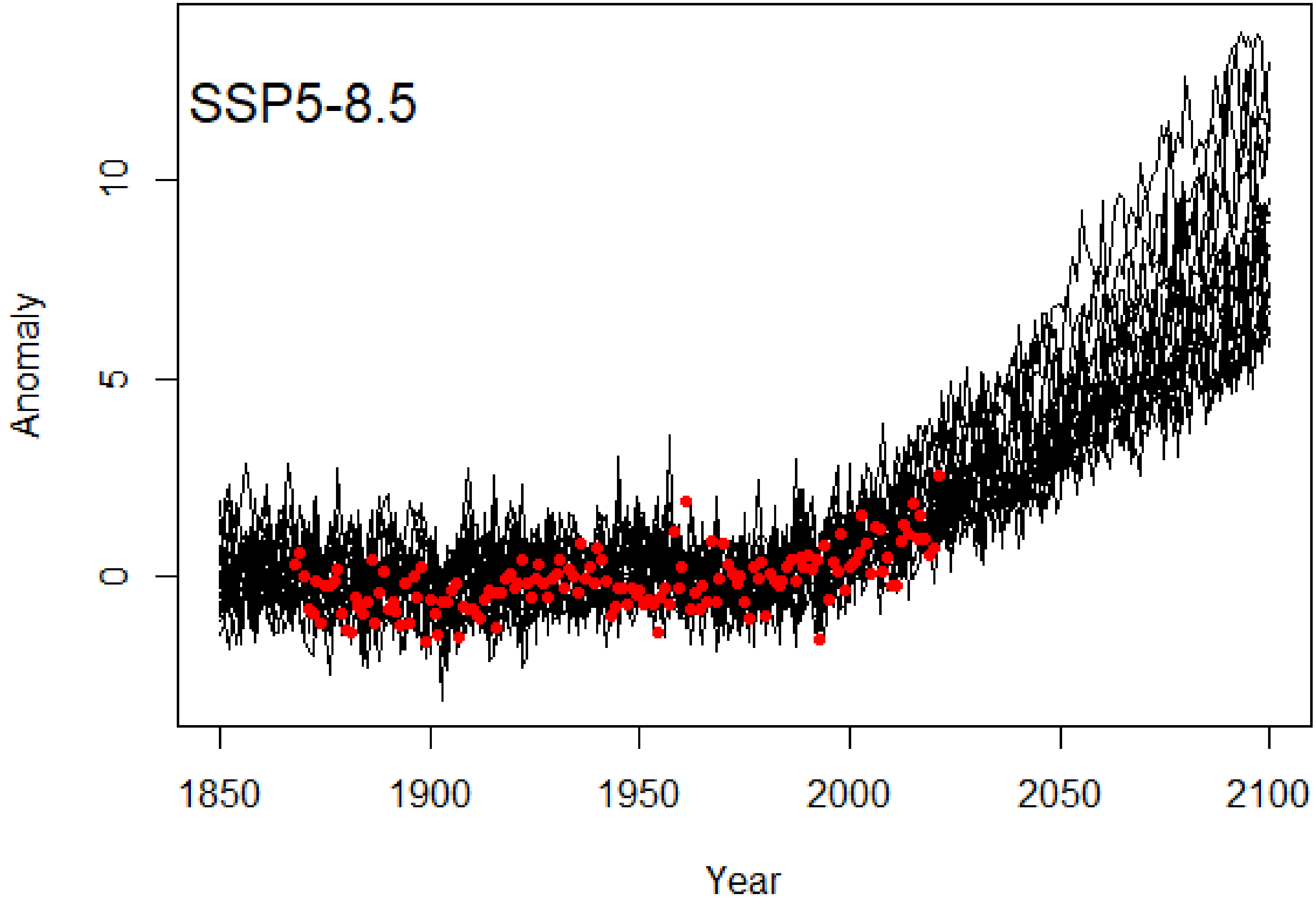
- We can also calculate the regional summer temperature over the same region from a climate model

- Use climate models to project forwards...

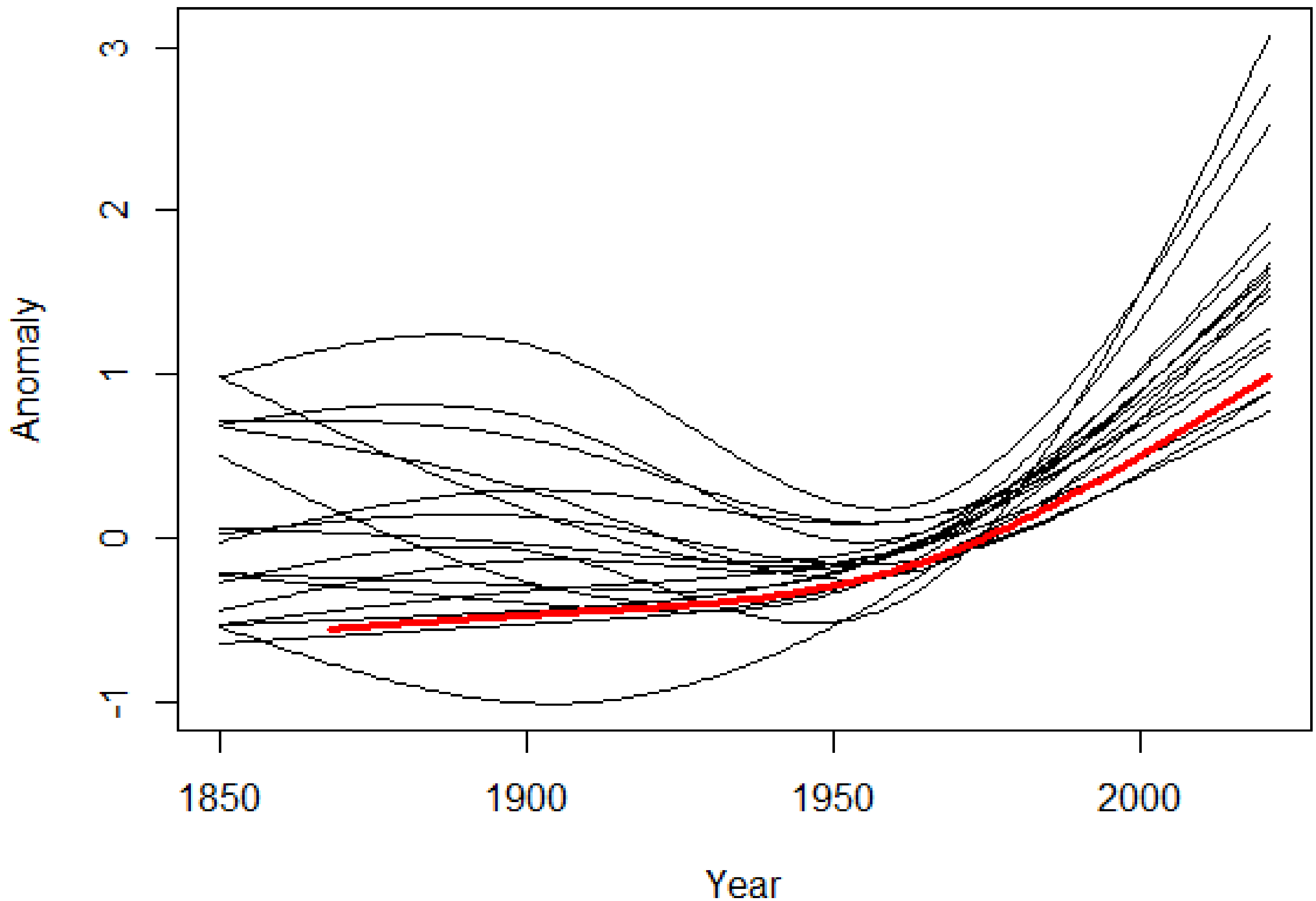
Regional Mean Tracks Annual Maxima Better Than Global Mean



Observed Regional Means and 17 Climate Models

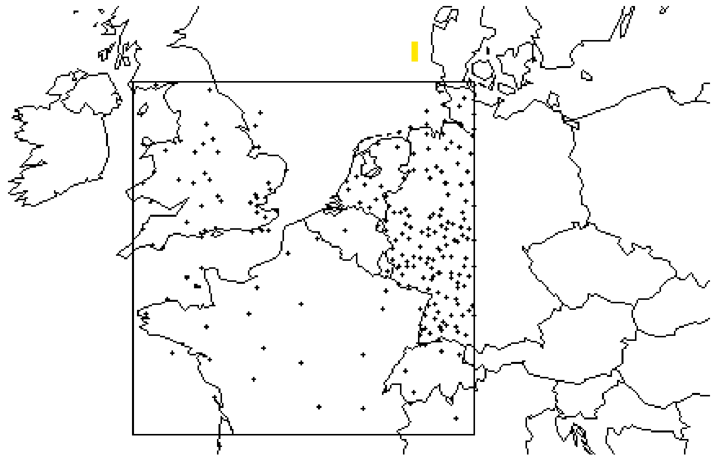


Models Overestimate the Regional Means
(Fitted spline surves with 3DF)

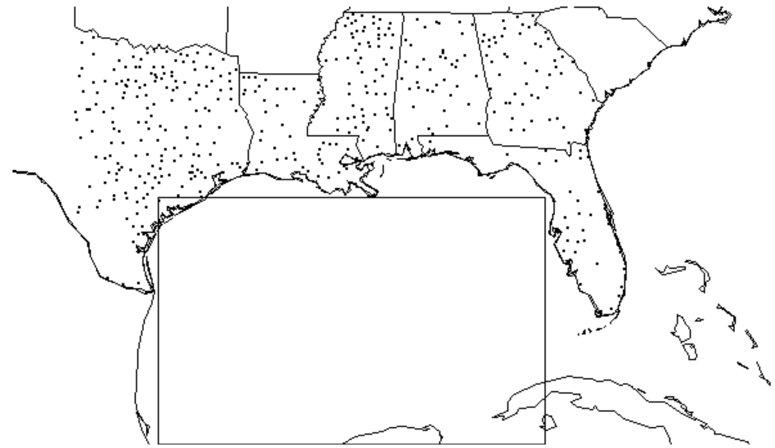


Similar Problems with Other Series

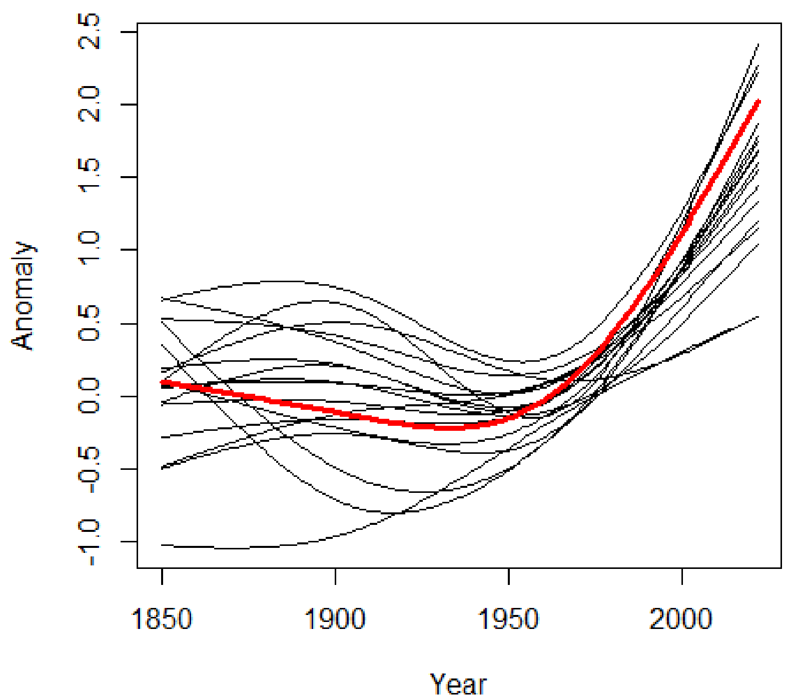
Boundaries for Northern Europe Region



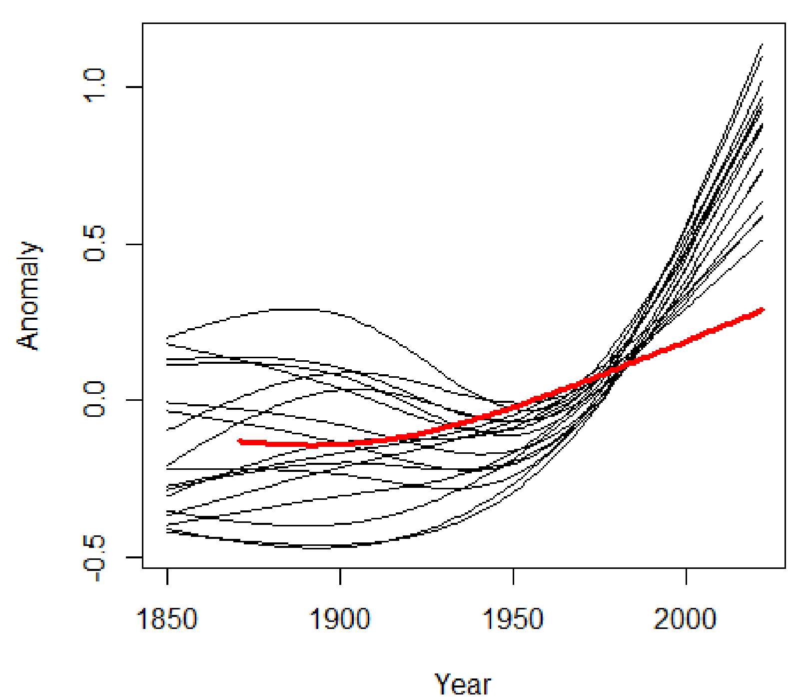
Boundaries for Gulf of Mexico Region



Splines for NEur Data



Splines for GoM Data



Precedents

- Meteorological:
  - PNW: Bercos-Hickey *et al.* (2022) — conditional hindcast attribution study; Fischer and Knutti (2023) — ensemble boosting approach;
  - Northern Europe: Vautard *et al.* (2023), Schumaker *et al.* (2024) — "Heat extremes in Western Europe increasing faster than simulated due to atmospheric circulation trends"
  - Gulf: Luo *et al.* (2016), Morey *et al.* (2017) — "SST increases in the western Gulf of Mexico...are...greatly reduced in the downscaled model compared to the low-resolution global model"
- Statistical:
  - Qasmi and Ribes (2022): "Kriging for climate change" (KCC) method
  - Katzfuss *et al.* (2017) — Bayesian approach to detection and attribution; not the same thing but related
  - Frühwirth-Schnatter *et al.* (2025), "Sparse Bayesian factor analysis when the number of factors is unknown"
  - Zhang *et al.* (2024) — spatial extremes model for PNW data

What We Want To Do

- Find a model that "corrects" for the discrepancy between climate models and observations, and also reflects disagreements among the climate models

- Use that model to project future values of the observations, given the model data

- Do this for the SSP5-8.5 emissions scenario, and also for two other well-known examples SSP2-4.5 and SSP1-2.6

Bayesian Factor Analysis Model
(following Bhattacharya—Dunson 2011)

Model response in year t with covariates $\{x_{tj},\ j=1,\ldots,J\}$ (splines with J df) and multiple signals $(h=1,\ldots,H)$:

$$u_{th} \;=\; \sum_{j=1}^{J} x_{tj}\gamma_{jh} + \eta_{th},\ \eta_{th} \sim \mathcal{N}(0, \phi_{\eta h}^{-1}\tau_h^{-1}),\ 1 \le t \le T,\ 1 \le h \le H, \quad (1)$$

Model outputs for $k=0,\ldots,K$ ($k=0$ are observations)

$$y_{tk} \;=\; \beta_{k0} + \sum_{h=1}^{H} \beta_{kh}u_{th} + \epsilon_{tk},\ \epsilon_{tk} \sim \mathcal{N}(0, \phi_{\epsilon k}^{-1}),\ k=0,\ldots,K, \quad (2)$$

Time periods $t \in [0,T]$ for climate models; $t \in [T_1, T_2]$ for observations

Priors:

$$\gamma_{jh} \;\sim\; \mathcal{N}[0, m_\gamma^{-1}],\ 1 \le j \le J,\ 1 \le h \le H, \quad (3)$$

$$\beta_{kh} \;\sim\; \mathcal{N}[0, m_\beta^{-1}],\ 0 \le k \le K,\ 0 \le h \le H, \quad (4)$$

$$\phi_{\eta h} \;\sim\; \mathrm{Gam}(a_\eta, b_\eta),\ 1 \le h \le H, \quad (5)$$

$$\phi_{\epsilon k} \;\sim\; \mathrm{Gam}(a_\epsilon, b_\epsilon,\ 0 \le k \le K), \quad (6)$$

$$\tau_h \;=\; \prod_{\ell=1}^{h} \delta_\ell,\ \delta_1 \sim \mathrm{Gam}(a_1, 1),\ \delta_\ell \sim \mathrm{Gam}(a_2, 1),\ \ell \ge 1,\ a_1 > 1,\ a_2 > 2. \quad (7)$$

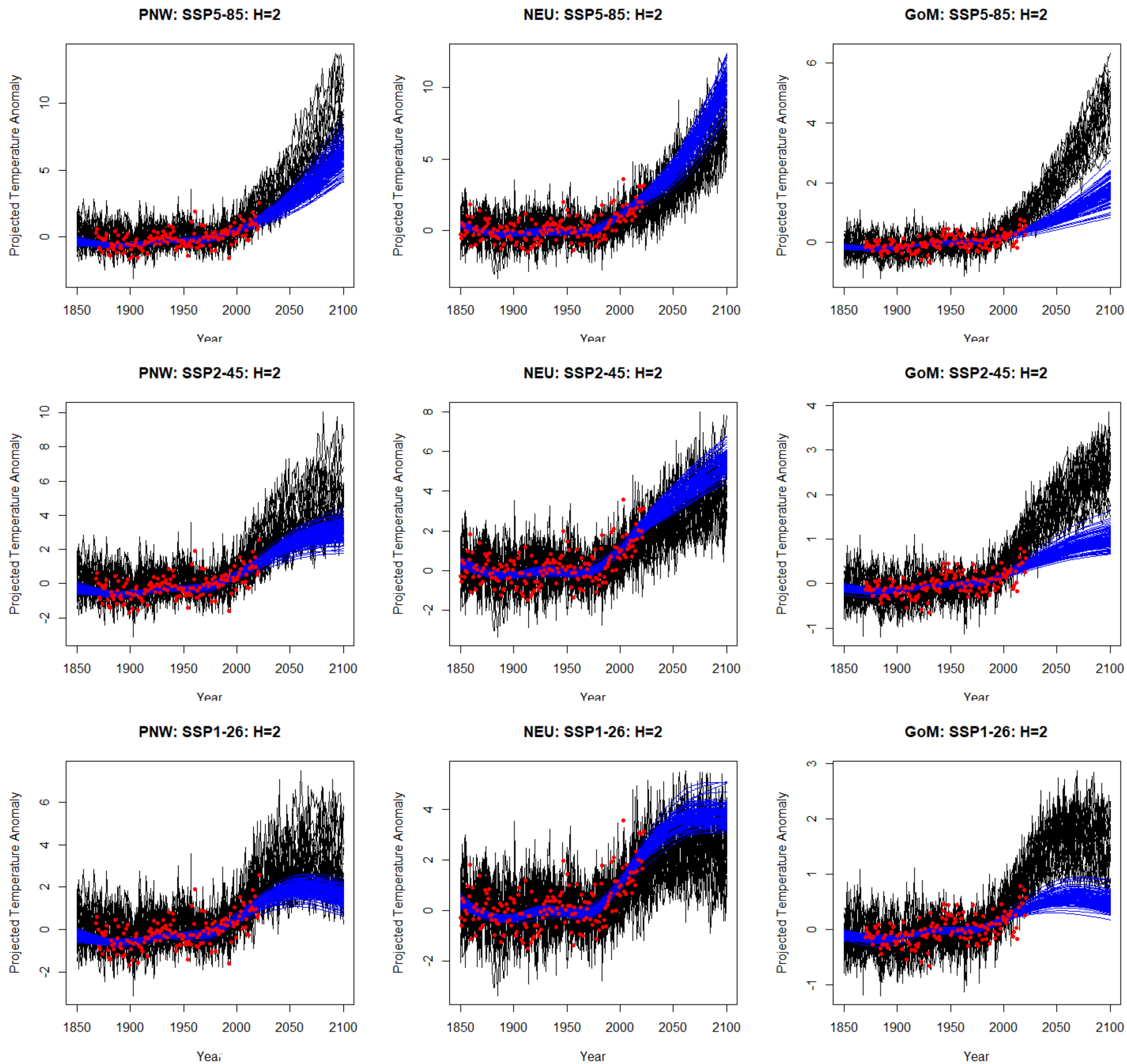
The τ_h's are shrinkage coefficients, designed to ensure convergence as $h \to \infty$.

Gibbs Sampler

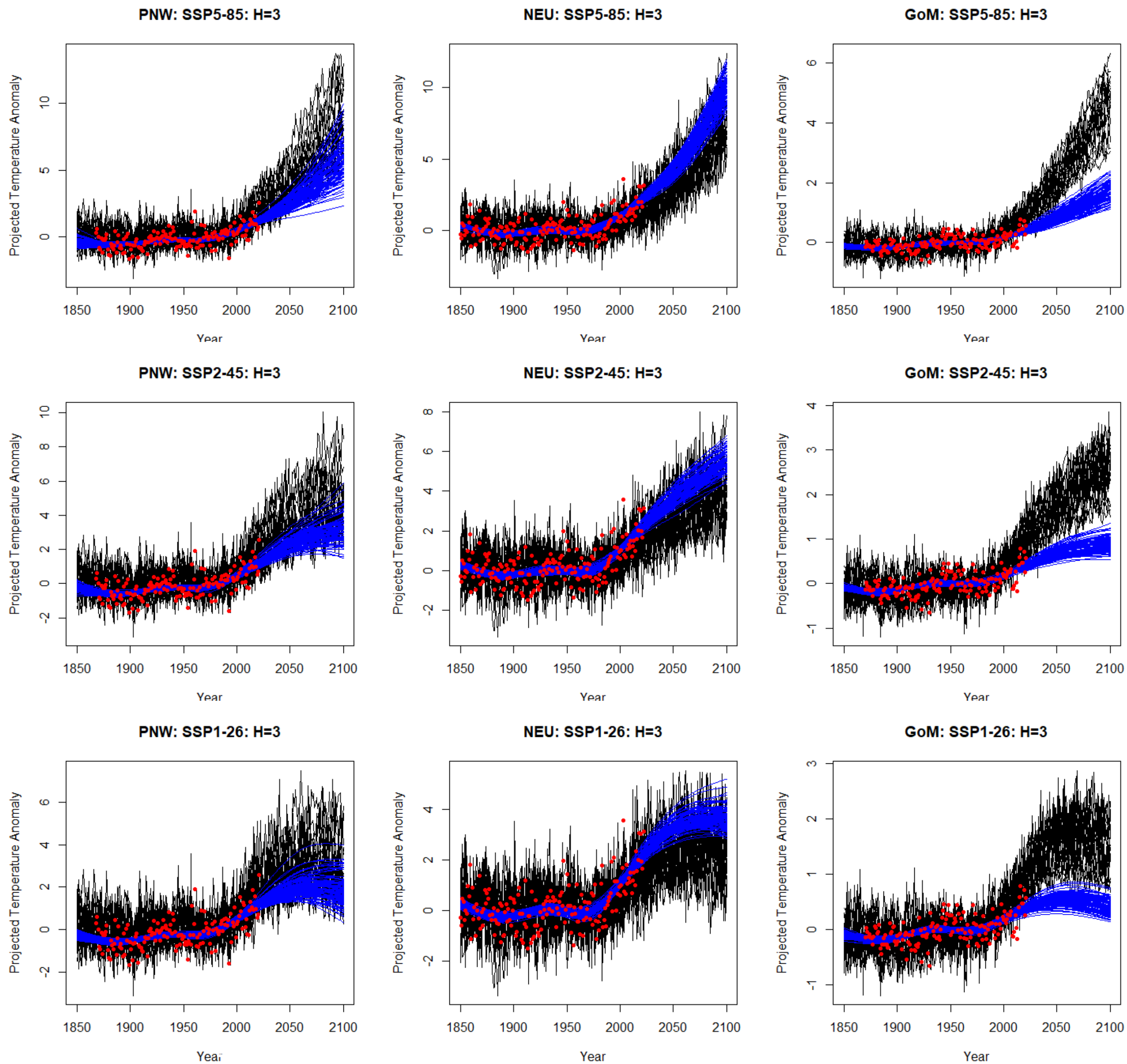
For each h define $\boldsymbol{\gamma}_h = \left(\gamma_{1h} \ \ldots \ \gamma_{Jh} \right)^T$, $\boldsymbol{\beta}_k = \left(\beta_{k0} \ \ldots \ \beta_{kH} \right)^T$, $\mathbf{u}_h = \left(u_{1h} \ \ldots \ u_{Th} \right)^T$, and X as the $T \times J$ matrix with entries x_{tj}. Conditional distributions for the Gibbs Sampler are given by

$$
\begin{aligned}
\boldsymbol{\gamma}_h|\ldots &\sim \mathcal{N}_J\left[\left(m_\gamma I_J + \phi_{\eta h}\delta_1\ldots\delta_h X^T X\right)^{-1}\phi_{\eta h}\delta_1\ldots\delta_h X^T\mathbf{u}_h, \left(m_\gamma I_J + \phi_{\eta h}\delta_1\ldots\delta_h X^T X\right)^{-1}\right] \\
\boldsymbol{\beta}_k|\ldots &\sim \mathcal{N}_{H+1}\left[\left(m_\beta I_{H+1} + \phi_{\epsilon_k} U_k^T U_k\right)^{-1}\phi_{\epsilon_k} U_k^T \mathbf{y}_k, \left(m_\beta I_{H+1} + \phi_{\epsilon_k} U_k^T U_k\right)^{-1}\right], \\
\mathbf{u}_h|\ldots &\sim \mathcal{N}_T\Bigg[\left(\phi_{\eta h}\delta_1\ldots\delta_h I_T + \sum_k \phi_{\epsilon k}\beta_{kh}^2 E_k^T E_k\right)^{-1}\left(\phi_{\eta h}\delta_1\ldots\delta_h X\boldsymbol{\gamma}_h + \sum_k \phi_{\epsilon k}\beta_{kh} E_k^T \tilde{\mathbf{y}}_{kh}\right), \\
&\qquad \left(\phi_{\eta h}\delta_1\ldots\delta_h I_T + \sum_k \phi_{\epsilon k}\beta_{kh}^2 E_k^T E_k\right)^{-1}\Bigg], \\
\phi_{\eta h}|\ldots &\sim \text{Gam}\left\{\frac{T}{2} + a_\eta, \frac{\delta_1\ldots\delta_h}{2}\left(\mathbf{u}_h - X\boldsymbol{\gamma}_h\right)^T\left(\mathbf{u}_h - X\boldsymbol{\gamma}_h\right) + b_\eta\right\}, \\
\delta_1|\ldots &\sim \text{Gam}\left\{\frac{TH}{2} + a_1, \frac{1}{2}\sum_h \phi_{\eta h}\left(\prod_{2\le\ell'\le h}\delta'_\ell\right)\left(\mathbf{u}_h - X\boldsymbol{\gamma}_h\right)^T\left(\mathbf{u}_h - X\boldsymbol{\gamma}_h\right) + 1\right\}, \\
\delta_\ell|\ldots &\sim \text{Gam}\left\{\frac{T(H-\ell+1)}{2} + a_2, \frac{1}{2}\sum_{h\ge\ell}\phi_{\eta h}\left(\prod_{1\le\ell'\le h,\ell'\ne\ell}\delta'_\ell\right)\left(\mathbf{u}_h - X\boldsymbol{\gamma}_h\right)^T\left(\mathbf{u}_h - X\boldsymbol{\gamma}_h\right)\right. \\
&\qquad \left. +1\right\}, \\
\phi_{\epsilon k}|\ldots &\sim \text{Gam}\left\{\frac{|\mathcal{T}_k|}{2} + a_\epsilon, \frac{1}{2}\left(\mathbf{y}_k - U_k\boldsymbol{\beta}_k\right)^T\left(\mathbf{y}_k - U_k\boldsymbol{\beta}_k\right) + b_\epsilon\right\}.
\end{aligned}
$$

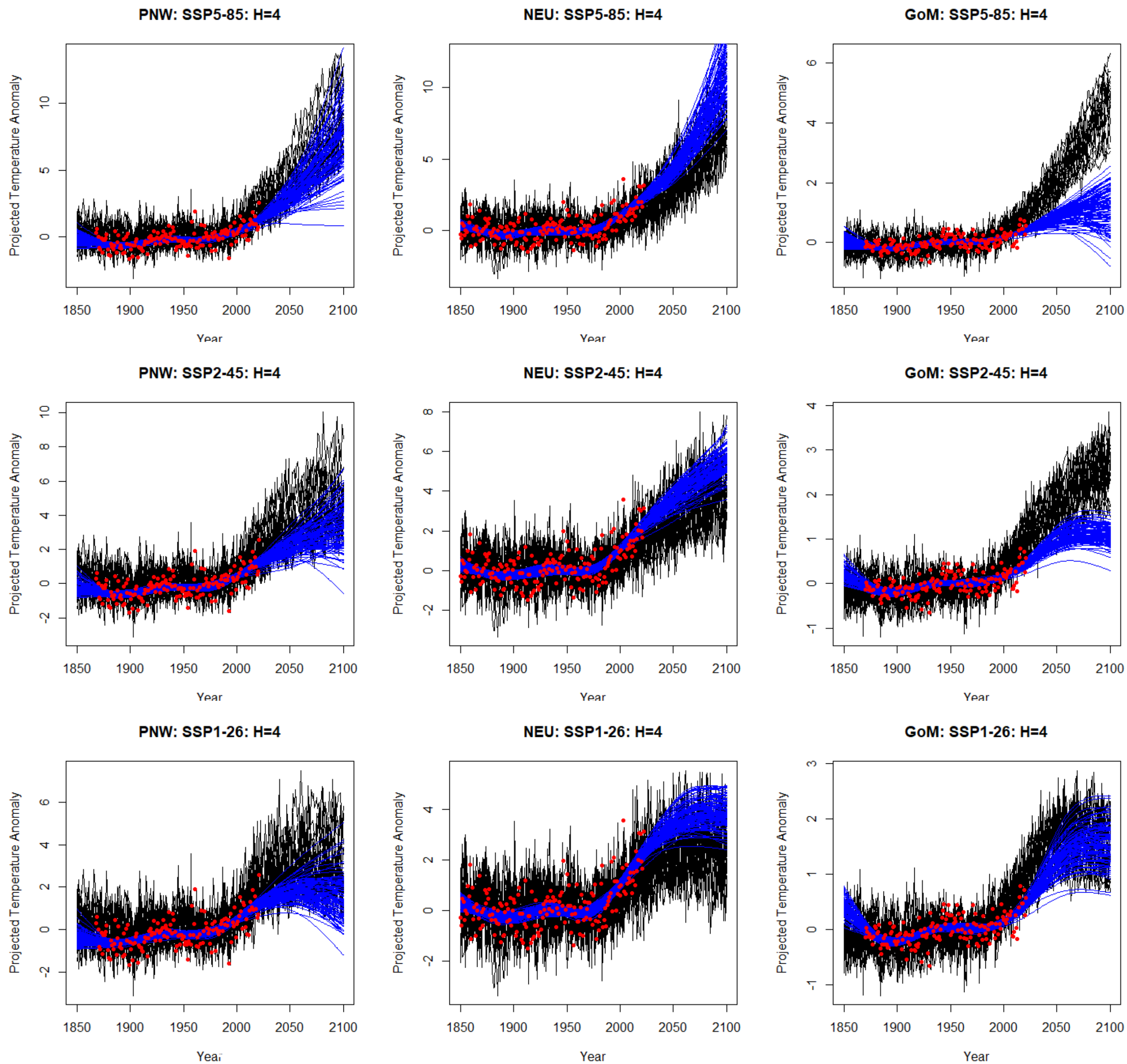
Results: Posterior Samples for Regional Means (H=2)



Results: Posterior Samples for Regional Means (H=3)



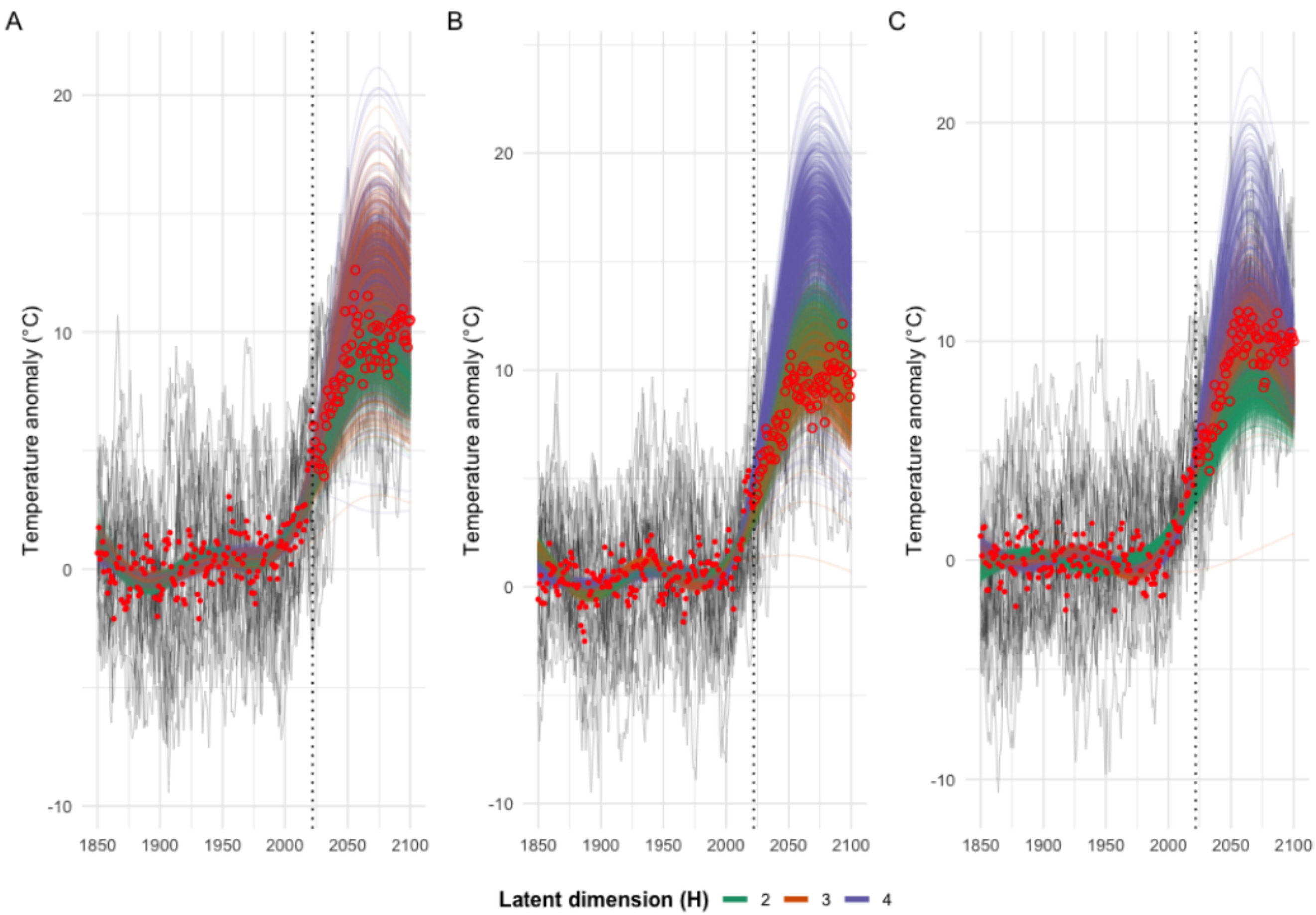
Results: Posterior Samples for Regional Means (H=4)



Methodological Issues

- Simulations of the underlying model
- Choice of H
  - DIC
  - Two ways of characterizing MSE (or RMSE)
- Which climate models should be excluded as outliers
- Number of iterations of MCMC sampler

Simulations for true H=2,3,4



Three methods of H selection

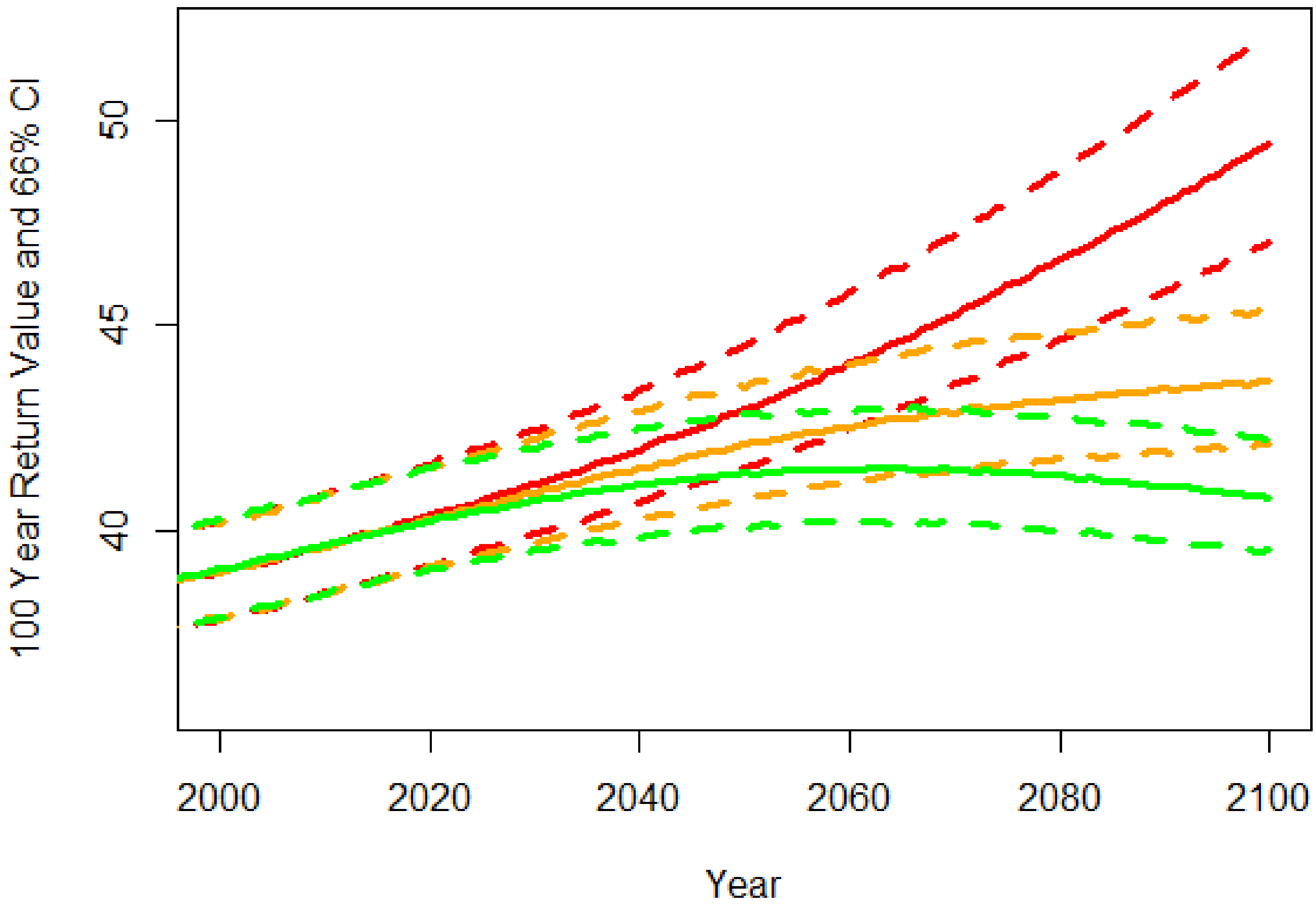
Table 1: Model comparison by true latent dimension (H_{true}) and fitted H. Lower values of DIC and RMSE indicate better fit and predictive performance.

True H	Fitted H	DIC	RMSE(masked)	RMSE(full)	RMSE(observed)
2	2	-1748	0.029	0.035	0.034
2	3	-1847	0.031	0.036	0.036
2	4	-1899	0.026	0.044	0.037
3	2	1673	0.185	0.120	0.086
3	3	-188	0.081	0.054	0.053
3	4	-350	0.101	0.062	0.055
4	2	3095	0.097	0.077	0.069
4	3	1112	0.069	0.061	0.053
4	4	-176	0.036	0.032	0.055

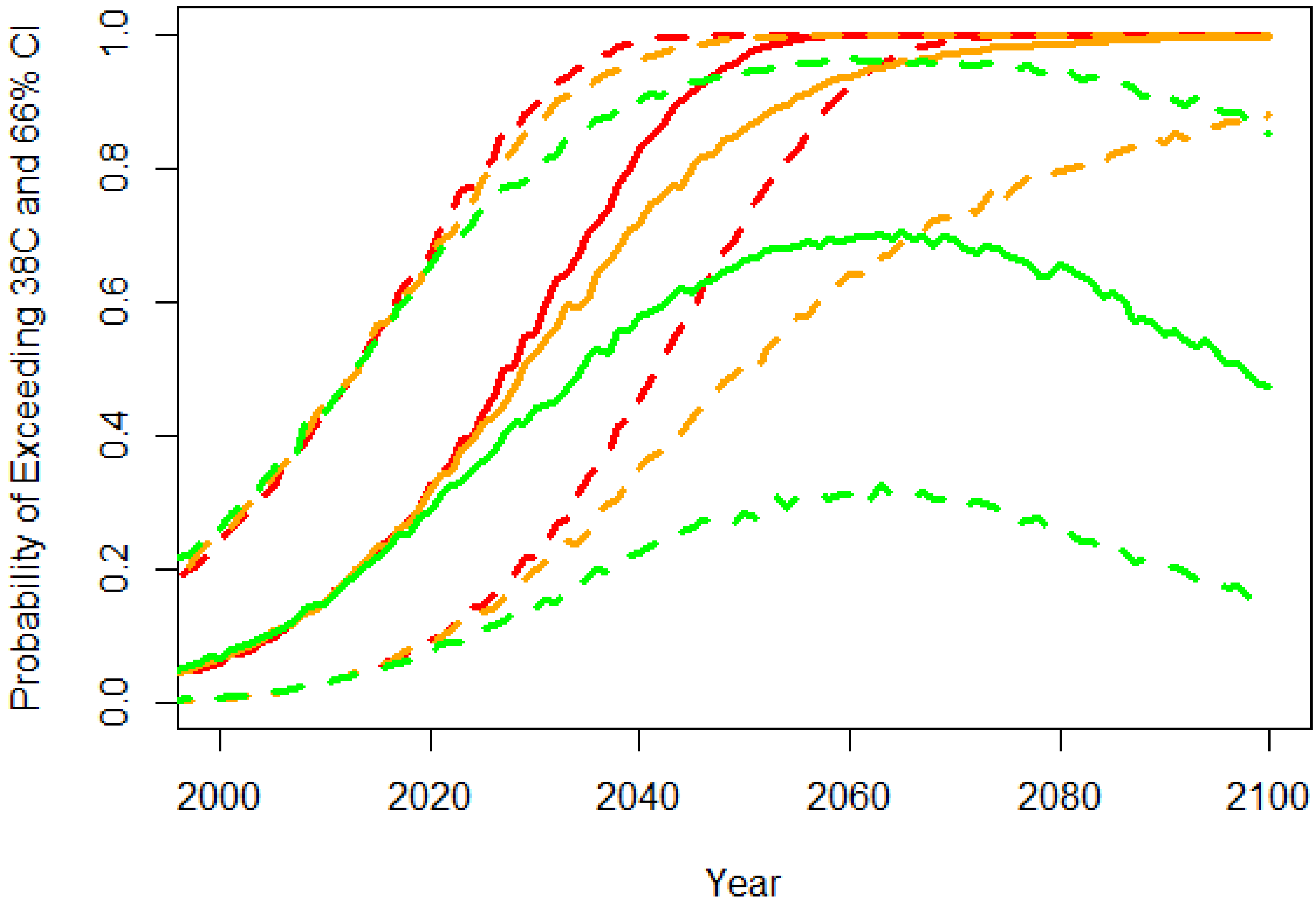
Application to Kelowna Extreme Temperature Projections

- Combine two Bayesian analyses:
  - GEV distribution for annual maxima with regional mean covariate, and
  - The analyses we've been describing for forward projections of regional means
- Repeat this analysis for three emission scenarios (SSP5-8.5, SSP2-4.5, SSP1-2.6)

Results for 100 Year Return Value



Results for Probability of Exceeding 38C



Discussion: Dynamical or Statistical Models?

- Sometimes, a debate arises whether methods like this are a substitute for better dynamical modeling

- My view: it's wrong to think of them as a competition

- Dynamical models are always preferable, but they need many more resources

- Even high-quality dynamical models have biases, and you still need statistical methods to process the outputs

- The methods proposed here may be viewed as a robust way to obtain projections based on pre-existing climate models, but they are not intended as a substitute for building better models

References

Bercos-Hickey E, O'Brien TA, Wehner MF, Zhang L, Patricola CM, Huang H and Risser MD (2022), Anthropogenic Contributions to the 2021 Pacific Northwest Heatwave, Geophysical Research Letters, 49, e2022GL099396. https://doi.org/10.1029/2022GL099396

Bhattacharya A and Dunson DB (2011), Sparse Bayesian infinite factor models. *Biometrika* 98, 291–306.

Frühwirth-Schnatter S, Hosszejni D and Lopes HF (2025), Sparse Bayesian Factor Analysis When the Number of Factors Is Unknown (with discussion), Bayesian Analysis 20(1), 213–314

Katzfuss M, Hammerling D and Smith RL (2017), A Bayesian hierarchical model for climate change detection and attribution. *Geophys. Res. Lett.* 44, 1–9, doi:10.1002/2017GL073688

Luo H, Bracco A, Cardona Y and McWilliams JC (2016), Submesoscale circulation in the Northern Gulf of Mexico: Surface processes and the impact of the freshwater river input. *Ocean Modeling*, DOI=10.1016/j.ocemod.2016.03.003

Morey S, Koch M, Liu Y, & Lee, S-K (2017), Florida's oceans and marine habitats in a changing climate. In E. P. Chassignet *et al.*, *Florida's climate: Changes, variations, & impacts*, pp. 391—425), https://doi.org/10.17125/fci2017.ch13

Qasmi S. and Ribes A. (2022), Reducing uncertainty in local temperature projections. *Sci. Adv.* 8, eabo6872, 12 October 2022.

Schumacher DL, Singh J, Hauser M, Fischer EM, Wild M and Seneviratne SI (2024), Exacerbated summer European warming not capturedby climate models neglecting long-term aerosol changes. Communications Earth & Environment 5:182

Vautard R, Cattiaux J,Happé T, Singh J, Bonnet R,Cassou C, Coumou D, D'Andrea F, Faranda D, Fischer E, Ribes A, Sippel S and Yiou P (2023), Heat extremes in Western Europe increasing faster than simulated due to atmospheric circulation trends. *Nature Communications* 14:6803

Zhang L, Risser MD, Wehner MF and O'Brien TA (2024), Leveraging extremal dependence to better characterize the 2021 Pacific Northwest heatwave. *JABES*, https://doi.org/10.1007/s13253-024-00636-8