Supplement to "Regional climate model assessment using statistical upscaling and downscaling techniques"

Veronica J. Berrocal[1,2], **Peter F. Craigmile**[3], **Peter Guttorp**[4,5]

[1] Department of Biostatistics, University of Michigan, Ann Arbor, MI 48109-2029

[2] Email: berrocal@umich.edu

[3] Department of Statistics, The Ohio State University, Columbus, OH 43210-1247, USA

[4] Department of Statistics, University of Washington, Seattle, WA 98195-4322, USA

[5] Norwegian Computing Center, Oslo, Norway

1 Synoptic station information

Table 1 provides some supplemental information on the seventeen SMHI synoptic stations used in the analysis.

2 Predictive process models

In this section we briefly review predictive process models [see also Banerjee et al., 2008] and we discuss how they are used in the statistical model for downscaling.

Let $Y(\mathbf{s})$ be a spatial process decomposed as:

$$Y(\mathbf{s}) = \mu(\mathbf{s}) + \eta(\mathbf{s}) + \epsilon(\mathbf{s}) \qquad \epsilon(\mathbf{s}) \overset{iid}{\sim} N(0, \tau^2), \tag{1}$$

Table 1: The seventeen SMHI synoptic stations used in the analysis. The end date is 2007-11-30 for all sites.

SMHI ID	Name	Start Date	Days Missing
2731	Sundsvalls Flygplats	1962-12-01	2
12226	Tånnås A	1995-11-23	0
12402	Sveg	1962-12-01	1
11743	Kuggören A	1995-12-17	0
10657	Åmot A	1995-08-01	12
10341	Malung	1962-12-01	0
10714	Films Kyrkby A	1982-02-02	129
9322	Karlstad Flygplats	1962-12-01	0
9720	Stockholm-Bromma	1962-12-01	14
9419	Kilsbergen-Suttarboda A	1995-08-01	0
9710	Tullinge A	1995-12-17	66
8226	Såtenås	1962-12-01	1
8524	Malmslått	1962-12-01	0
8714	Harstena A	1996-04-02	0
7446	Jönköpings Flygplats	1965-03-02	0
7840	Visby Flygplats	1962-12-01	0
7524	Målilla	1965-01-02	61

where $\mu(\mathbf{s})$ denotes the mean of $Y(\mathbf{s})$, $\eta(\mathbf{s})$ is a mean-zero Gaussian process and $\epsilon(\mathbf{s})$ is independent and identically distributed (IID) Gaussian noise. If the spatial dependence in $\eta(\mathbf{s})$ is modeled using a covariance function $C(\cdot,\cdot;\boldsymbol{\theta})$, inferring upon $\boldsymbol{\theta}$ is computationally challenging if the number of sites with observations is large. The predictive process modeling framework alleviates the computational burden by projecting $\eta(\mathbf{s})$ onto a lower dimensional space spanned by a realization of $\eta(\mathbf{s})$ at selected sites.

Specifically, let $\mathbf{r}_1^\star,\ldots,\mathbf{r}_m^\star$ be a set of m knots and let $\boldsymbol{\eta}^\star$ be the $m\times 1$ random vector $(\eta(\mathbf{r}_1^\star),\ldots,\eta(\mathbf{r}_m^\star))'$. Then, $\boldsymbol{\eta}^\star$ has a multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix $\mathbf{C}^\star(\boldsymbol{\theta})$ induced by the covariance function $C(\cdot,\cdot;\boldsymbol{\theta})$.

The predictive process $\tilde{\eta}(\mathbf{s})$ derived by the parent process $\eta(\mathbf{s})$ is defined at each site $\mathbf{s}$ via the following kriging equation:

$$\tilde{\eta}(\mathbf{s}) = c'(\mathbf{s};\boldsymbol{\theta})\cdot \mathbf{C}^{\star^{-1}}(\boldsymbol{\theta})\cdot\boldsymbol{\eta}^\star \tag{2}$$

where $c(\mathbf{s};\boldsymbol{\theta})$ is a $m\times 1$ vector with i-th component equal to $C(\mathbf{s},\mathbf{r}_i^\star;\boldsymbol{\theta})$. It is clear then that inferring upon $\tilde{\eta}(\mathbf{s})$ is less cumbersome than inferring upon the parent process $\eta(\mathbf{s})$ since it only requires inverting an $m\times m$ covariance matrix.

In our statistical model for downscaling we use predictive processes to alleviate the computational burden associated with inferring upon the spatial process $Q(\mathbf{r},t)$ that define the spatially and temporally varying weights $w_k(\mathbf{s},t)$. Thus, we select m=308 knots by systematically subsampling the set of g=2,640 climate model grid boxes centroids and we define the weights by replacing $Q(\mathbf{r},t)$ with the predictive process $\tilde{Q}(\mathbf{r},t)$.

3 Downscaling model fitting details

We fit the downscaling model presented in Section 3 using a Markov Chain Monte Carlo (MCMC) algorithm. The priors placed on the model parameters are all conjugate priors, and thus all the parameters, with the exception of the quarter-specific decay parameter $\phi_{0,t}$ and the latent spatio-temporal predictive process $\tilde{Q}(\mathbf{r}, t)$ are updated using a Gibbs sampling algorithm. To update the decay parameters $\phi_{0,t}$, we use a Metropolis-Hastings algorithm: in each MCMC iteration, for each $t = 1, \ldots, T$, we sample a candidate value $\phi'_{0,t}$ from a discrete uniform distribution placed on a grid of 200 values ranging from 0.005 to 0.5. We compute the Metropolis-Hastings ratio and we accept or reject the proposed value according to the usual Metropolis-Hastings scheme.

To update the predictive process $\tilde{Q}(\mathbf{r}, t)$, $t = 1, \ldots, T$ at each MCMC iteration, we simply need to update, for each $t = 1, \ldots, T$, the m-dimensional (m=308) random vector $\mathbf{Q}^\star$, realization of the parent process $Q(\mathbf{r}, t)$. We achieve this by proposing at each iteration a candidate value $\mathbf{Q}^{\star'}$ from an m-dimensional multivariate normal distribution centered around the current value of $\mathbf{Q}^\star$ and with a diagonal covariance matrix with standard deviation chosen so to have an acceptance rate of 25-40%. The proposed value is then accepted or rejected according to the usual Metropolis-Hastings scheme and the corresponding predictive process $\tilde{Q}(\mathbf{r}, t)$ is derived according to (2).

4 Downscaling model predictions over grid boxes

In this section we briefly discuss the approach undertaken to generate predictions over grid boxes under the downscaling model framework. Let B_v, $v = 1, \ldots, g$, denote a climate model grid box, and let i denote the i-th MCMC iteration. To generate predictions over grid boxes we sample from the posterior predictive distribution. Thus, at the i-th MCMC iteration we predict the annual average temperature $Y(B_v, t)$ over grid box B_v via:

$$Y^{(i)}(B_v, t) = \frac{1}{|B_v|} \int_{B_v} \tilde{\beta}_{0,t}^{(i)}(\mathbf{s}, t) + \beta_1^{(i)} \tilde{x}^{(i)}(\mathbf{s}, t) + \epsilon^{(i)}(\mathbf{s}, t) d\mathbf{s}, \tag{3}$$

where the superscript denotes the value of the parameter at the i-th MCMC iteration.

We compute the integral in (3) via numerical integration by taking a systematic sample of q sites $\mathbf{s}_1, \ldots, \mathbf{s}_q$ within grid cell B_v replacing (3) with

$$\overline{Y}^{(i)}(B_v, t) = \frac{1}{q} \sum_{j=1}^{q} \left[\tilde{\beta}_0^{(i)}(\mathbf{s}_j, t) + \beta_1^{(i)} \tilde{x}^{(i)}(\mathbf{s}_j, t) + \epsilon^{(i)}(\mathbf{s}_j, t) \right]. \tag{4}$$

Given the large number of climate model grid boxes, g, computing (4) is computationally challenging within the downscaling model framework. Thus, we further approximate (4) with the prediction at the centroid of the grid box, $\mathbf{r}_v$ say:

$$Y^{(i)}(\mathbf{r}_v, t) = \tilde{\beta}_0^{(i)}(\mathbf{r}_v, t) + \beta_1^{(i)} \tilde{x}^{(i)}(\mathbf{r}_v, t) + \epsilon^{(i)}(\mathbf{r}_v, t). \tag{5}$$

This approximation of $\overline{Y}^{(i)}(B_v, t)$ with $Y^{(i)}(\mathbf{r}_v, t)$ is valid only in expectation, that is:

$$E\left[\overline{Y}(B_v, t) | \text{data}\right] \approx E\left[Y(\mathbf{r}_v, t) | \text{data}\right].$$

We now show that such an approximation is warranted within the downscaling model frame-

work. [1] Since $\tilde{\beta}_0(\mathbf{s},t)$ is decomposed as $\beta_{0,t} + \beta_0(\mathbf{s},t)$ with $\beta_0(\mathbf{s},t)$ mean-zero Gaussian process:

$$\frac{1}{q}\sum_{j=1}^{q}\tilde{\beta}_0^{(i)}(\mathbf{s}_j,t) = \beta_{0,t}^{(i)} + \frac{1}{q}\sum_{j=1}^{q}\beta_0^{(i)}(\mathbf{s}_j,t)$$

On the other hand, $\tilde{\beta}_0^{(i)}(\mathbf{r}_v,t) = \beta_{0,t}^{(i)} + \beta_0^{(i)}(\mathbf{r}_v,t)$. Averaging across MCMC iterations warrants that we can approximate $\frac{1}{q}\sum_{j=1}^{q}\tilde{\beta}_0^{(i)}(\mathbf{s}_j,t)$ with $\tilde{\beta}_0^{(i)}(\mathbf{r}_v,t)$. Similar arguments can be used to justify approximating $\frac{1}{q}\sum_{j=1}^{q}\epsilon^{(i)}(\mathbf{s}_j,t)$ with $\epsilon^{(i)}(\mathbf{r}_v,t)$, in expectation.

We are now left to show that $\frac{1}{q}\sum_{j=1}^{q}\beta_1^{(i)}\tilde{x}^{(i)}(\mathbf{s}_j,t)$ can be approximated with $\beta_1^{(i)}\tilde{x}^{(i)}(\mathbf{r}_v,t)$. For each $\mathbf{s}$ and $t = 1,\ldots,T$, the smoothed climate model output $\tilde{x}(\mathbf{s},t)$ is defined as

$$\tilde{x}(\mathbf{s},t) = \sum_{k=1}^{g} w_k(\mathbf{s},t)x(B_k,t),$$

where $x(B_k,t)$ is the climate model output at grid box B_k for quarter t and $w_k(\mathbf{s},t)$ are the site-specific weights defined as:

$$w_k(\mathbf{s},t) = \frac{\exp(-\lambda|\mathbf{s}-\mathbf{r}_k|)\cdot\exp(Q(\mathbf{r}_k,t))}{\sum_{l=1}^{g}\exp(-\lambda|\mathbf{s}-\mathbf{r}_l|)\cdot\exp(Q(\mathbf{r}_l,t))}. \tag{6}$$

Thus, to show that $\frac{1}{q}\sum_{j=1}^{q}\beta_1^{(i)}\tilde{x}^{(i)}(\mathbf{s}_j,t)$ can be approximated with $\beta_1^{(i)}\tilde{x}^{(i)}(\mathbf{r}_v,t)$ is sufficient to show that $\frac{1}{q}\sum_{j=1}^{q}\tilde{x}^{(i)}(\mathbf{s}_j,t) \approx \tilde{x}^{(i)}(\mathbf{r}_v,t)$, or equivalently that

$$\frac{1}{q}\sum_{j=1}^{q} w_k(\mathbf{s}_j,t) \approx w_k(\mathbf{r}_v,t) \qquad k = 1,\ldots,g. \tag{7}$$

Since the q points $\mathbf{s}_j$ are a systematic sample in the 12.5 km×12.5 km grid box B_v with centroid $\mathbf{r}_v$ and $\lambda = 0.08$, it follows that for $k = 1,\ldots,g$

$$\exp(-\lambda|\mathbf{s}_j-\mathbf{r}_k|) \approx \exp(-\lambda|\mathbf{r}_v-\mathbf{r}_k|).$$

[1] We remark here that the same numerical integration method is used for producing grid box predictions with the upscaling model, but that no approximation of the average is used for that model.

Hence (7) holds and $\frac{1}{q}\sum_{j=1}^{q}\beta_1^{(i)}\tilde{x}^{(i)}(\mathbf{s}_j,t)$ can be approximated with $\beta_1^{(i)}\tilde{x}^{(i)}(\mathbf{r}_v,t)$.

Further research will be needed to quantify the difference in the variability of $\overline{Y}^{(i)}(B_v,t)$ and $Y^{(i)}(\mathbf{r}_v,t)$.

5 Posterior summaries for the downscaling model

In this section we present estimates of the model parameters for our downscaling model. Figure 1 shows the posterior mean of the additive calibration term $\beta_{0,t}$ along with its 95% credible interval for each quarter during the study period 1962-2007. The additive term $\beta_{0,t}$ provides information on the additive bias of the regional climate model output. Spatial adjustment to the overall calibration term $\beta_{0,t}$ is accounted for by $\beta_0(\mathbf{s},t)$. Figure 2 displays the posterior predictive mean of $\beta_0(\mathbf{s},t)$ for each quarter in year 2002 over the entire spatial domain while Figure 3 presents the posterior predictive standard deviation of $\beta_0(\mathbf{s},t)$ for each quarter in 2002.

Finally, Table 2 presents the posterior mean along with the 95% credible interval for all the time invarying model parameters: the multiplicative bias term β_1; the variance σ_Q^2 of the latent spatio-temporal process $Q(\mathbf{s},t)$ used to construct the spatio-temporal set of weights $w_k(\mathbf{s},t)$; the variance $\sigma_{\beta_0}^2$ of the spatially-varying additive bias $\beta_0(\mathbf{s},t)$; and the residual variance τ^2 in the seasonal temperature data that cannot be accounted for by the smoothed regional climate model output.

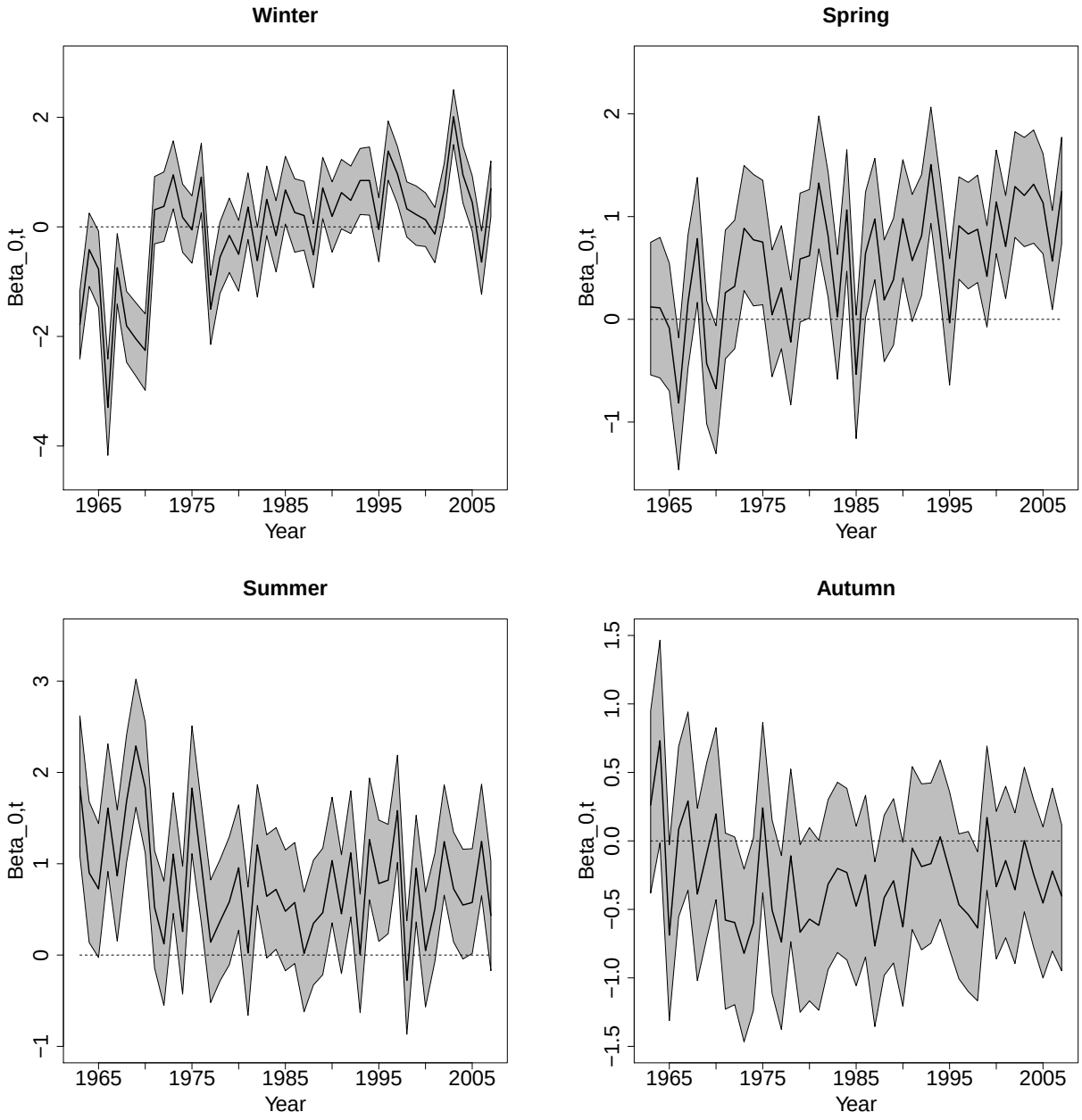

Figure 1: Posterior mean of the additive calibration term $\beta_{0,t}$ (black solid line) and 95% credible interval (gray bands) for each quarter during between December 1, 1962 and November 30, 2007.

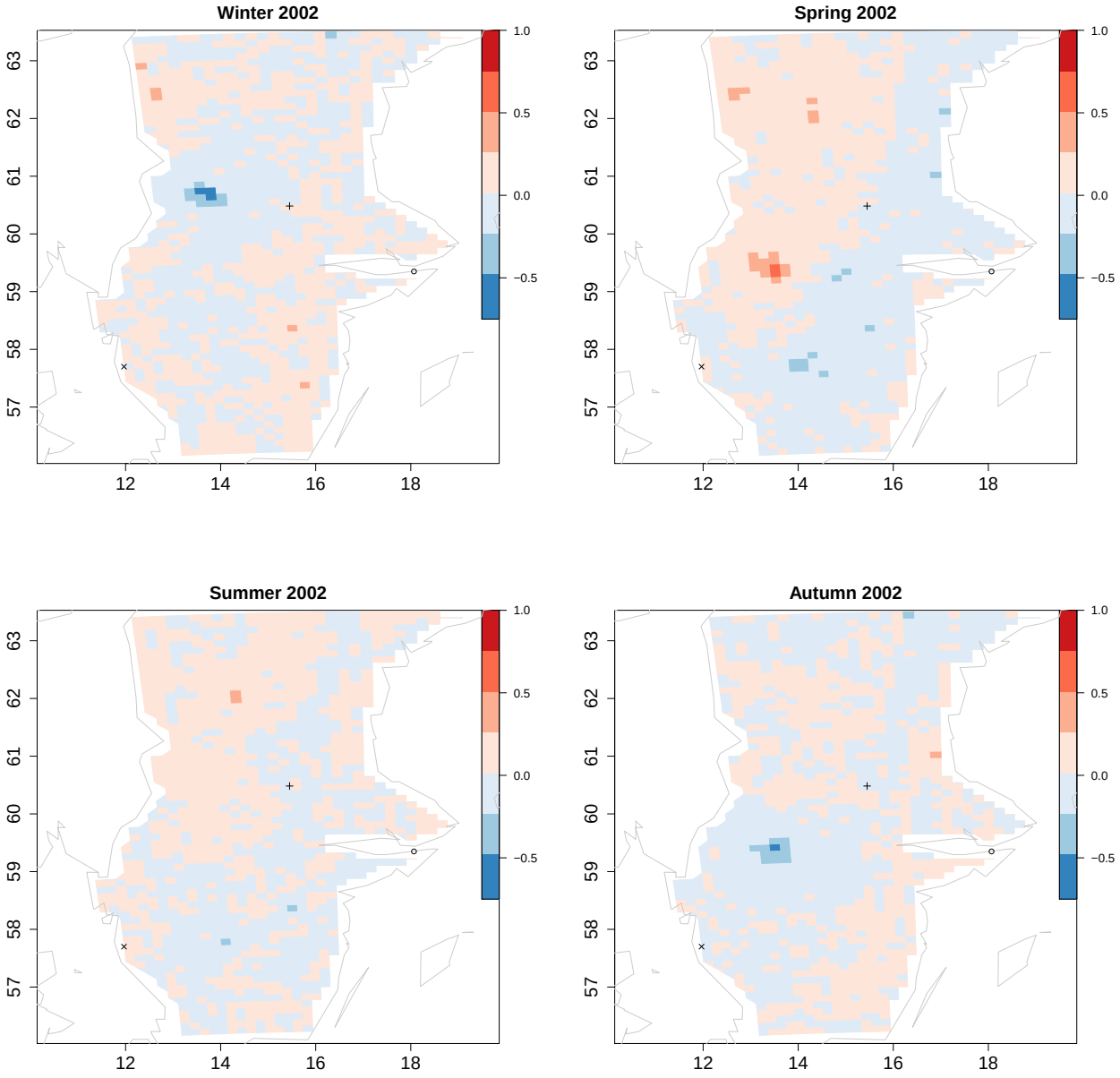

Figure 2: Posterior predictive mean of the spatially varying additive calibration term $\beta_0(\mathbf{s}, t)$ for each quarter in year 2002.

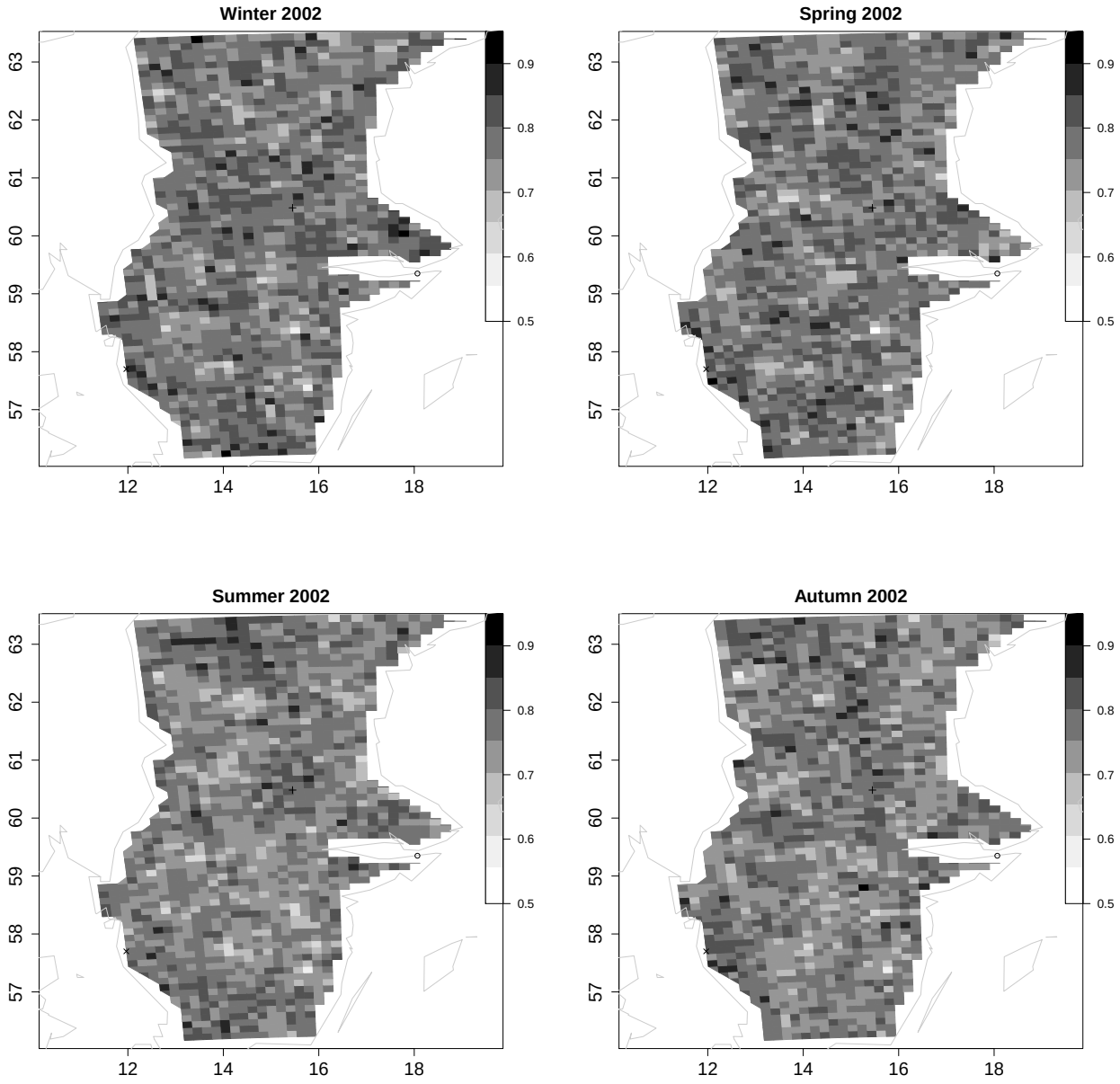

Figure 3: Posterior predictive standard deviation of the spatially varying additive calibration term $\beta_0(\mathbf{s}, t)$ for each quarter in year 2002.

Table 2: Posterior means and 95% credible intervals for all the parameters of the downscaling statistical model that are constant in time.

Parameter	Posterior mean	95% credible interval
β_1	0.93	(0.91, 0.95)
σ^2_Q	1.33	(1.18, 1.50)
$\sigma^2_{\beta_0}$	0.61	(0.43, 0.80)
τ^2	0.31	(0.17, 0.46)

6 Posterior summaries for the upscaling model

Tables 3, 4, and 5 display numerical posterior summaries of a subset of the trend, seasonal, and noise hyperparameters in the upscaling model. Graphical summaries of the daily posterior distributions of the trend component $\{\mu(\boldsymbol{s}, t\}$ is shown in Figure 4 and the seasonal component $\{\psi(\boldsymbol{s}, t\}$ is show in Figure 5. Figures 6 and 7 show, respectively, posterior summaries of the parameters characterizing the short and long range dependence in the noise, and the parameters of the space-time volatility term.

7 Comparison of predicted seasonal fields

Figure 8 presents the average difference between the average quarterly temperature fields predicted by the downscaling model over the entire region and the average quarterly temperature predicted by the RCM for each quarter in the study period December 1, 1962-

Trend	Parameter	Mean	95% credible interval
$V_\tau(\boldsymbol{s})$	mean	117.45	(106.39, 128)
	AR(1) par.	0.42	(0.02, 0.83)
	sill	1452.99	(540.39, 2264.46)
	range	7.5	(5.09, 10.72)

Table 3: Posterior summaries (means and 95% posterior credible intervals) for the trend hyperparameters in the model.

November 30, 2007, and the analogous difference between the average quarterly temperature fields predicted by the downscaling model over the entire region and the average quarterly temperature predicted by the RCM. In both cases, the average difference between the predictions has been computed over the RCM grid cells that cover the spatial domain.

References

A. Banerjee, A. E. Gelfand, A. O. Finley, and H. Sang. Gaussian predictive process models for large spatial data sets. *Journal of the Royal Statistical Society Series B*, 70:825–848, 2008.

Seasonal	Parameter	Mean	95% credible interval
$\gamma_1(\boldsymbol{s})$	mean	-3.49	(-4.56, -2.43)
	sill	0.38	(0.06, 1.16)
	range	9.93	(4.80, 17.07)
$\gamma_2(\boldsymbol{s})$	mean	-8.92	(-12.95, -4.61)
	sill	7.17	(2.36, 18.27)
	range	9.81	(4.67, 17.06)
$\gamma_3(\boldsymbol{s})$	mean	0.24	(-0.68, 1.21)
	sill	0.3	(0.03, 1.07)
	range	9.86	(4.68, 16.92)
$\gamma_4(\boldsymbol{s})$	mean	0.62	(0.07, 1.17)
	sill	0.10	(0.02, 0.33)
	range	10.22	(5.19, 17.31)

Table 4: Posterior summaries (means and 95% posterior credible intervals) for the seasonal hyperparameters in the model.

Noise	Parameter	Mean	95% credible interval
$\kappa(\boldsymbol{s})$	mean	1.28	(0.87, 1.68)
	sill	0.06	(0.01, 0.19)
	range	10.25	(5.08, 16.98)
$\theta_1(\boldsymbol{s})$	mean	0.97	(0.65, 1.31)
	sill	0.04	(0.01, 0.11)
	range	10.05	(4.8, 17.13)
$\theta_2(\boldsymbol{s})$	mean	0.28	(0.12, 0.42)
	sill	0.01	(0.00, 0.02)
	range	11.7	(6.05, 19.46)

Table 5: Posterior summaries (means and 95% posterior credible intervals) for the noise hyperparameters in the model.

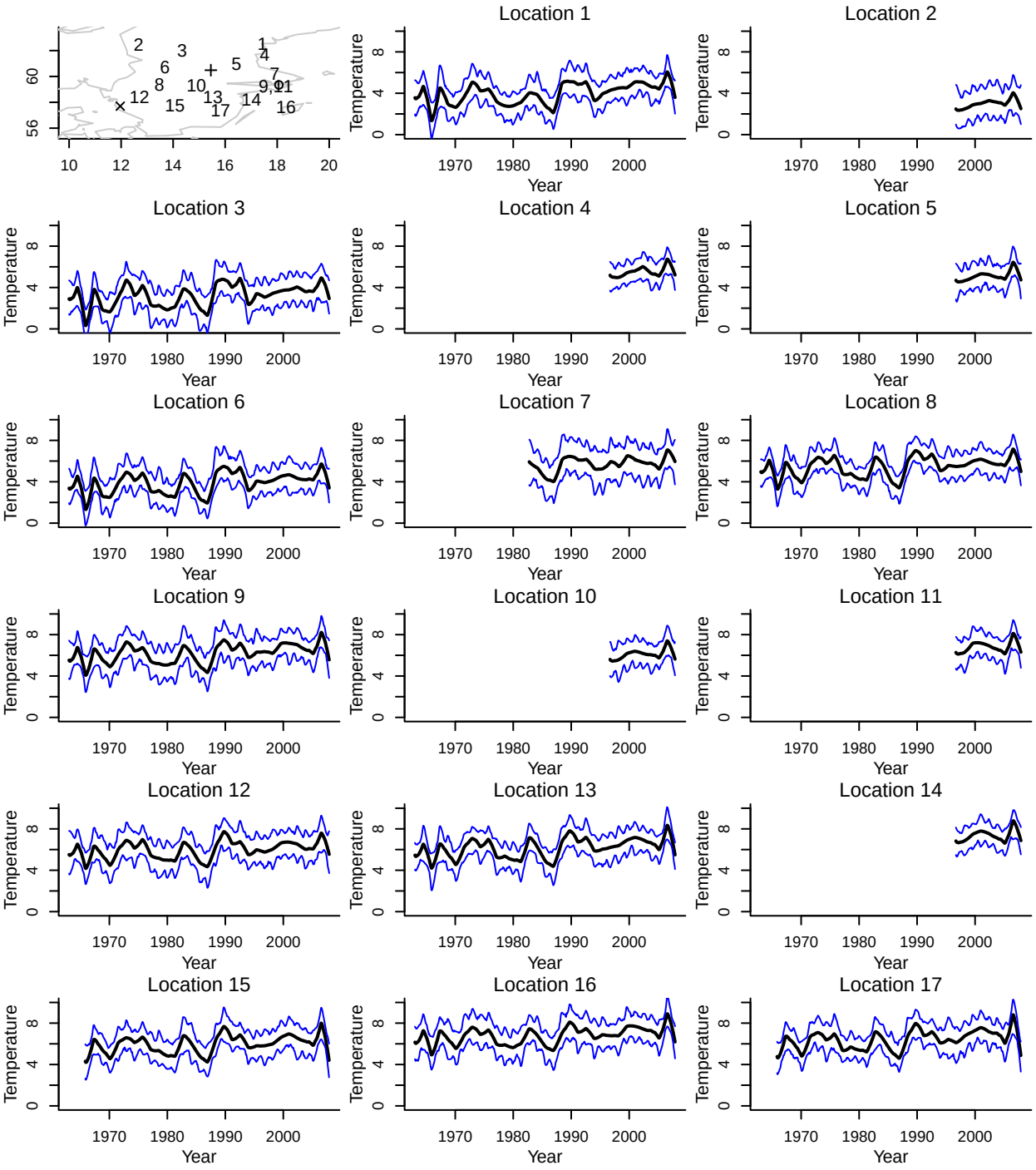

Figure 4: By location, the daily posterior mean of the trend $\{\mu(\boldsymbol{s}, t)\}$ at each observed location (black line), as well as 95% credible intervals for the parameter (blue lines).

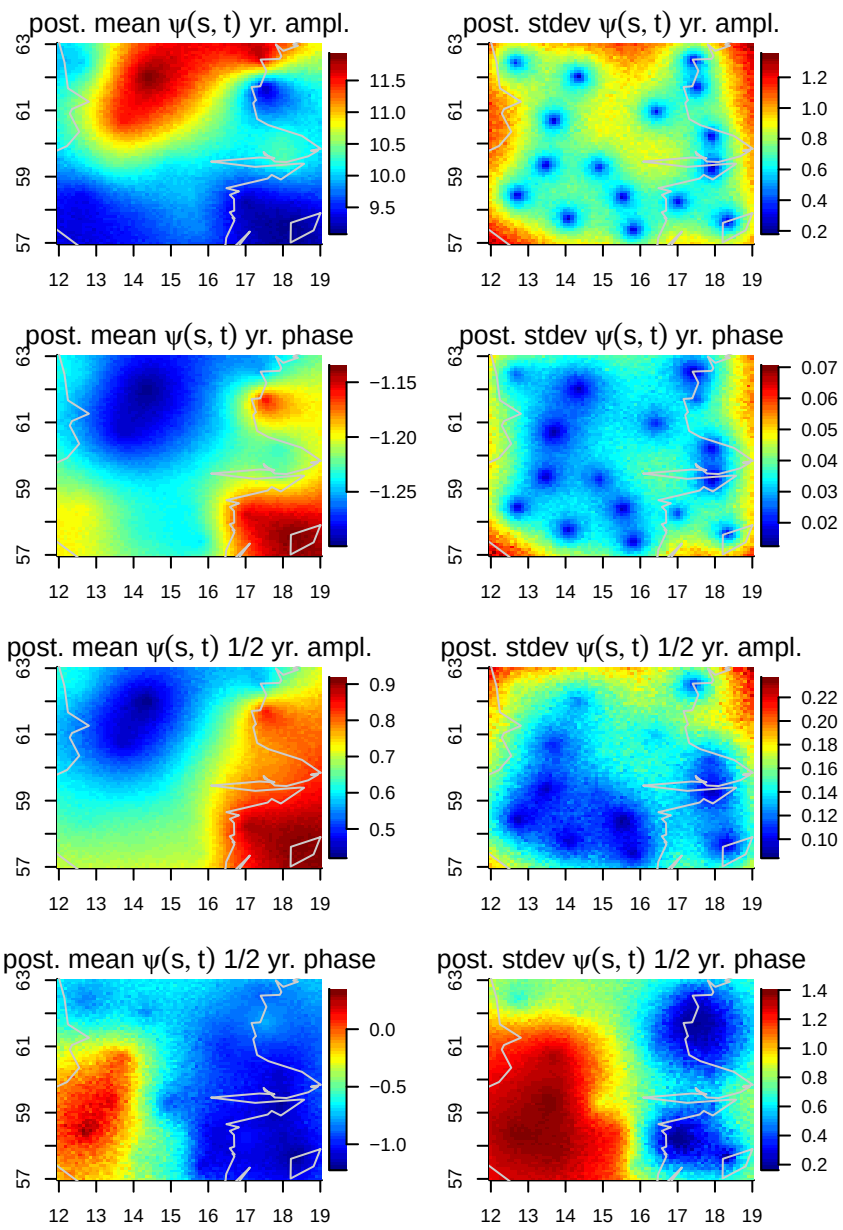

Figure 5: Posterior mean fields (left panels) and standard deviation fields (right panels) for the parameters characterizing the seasonality in daily temperatures.

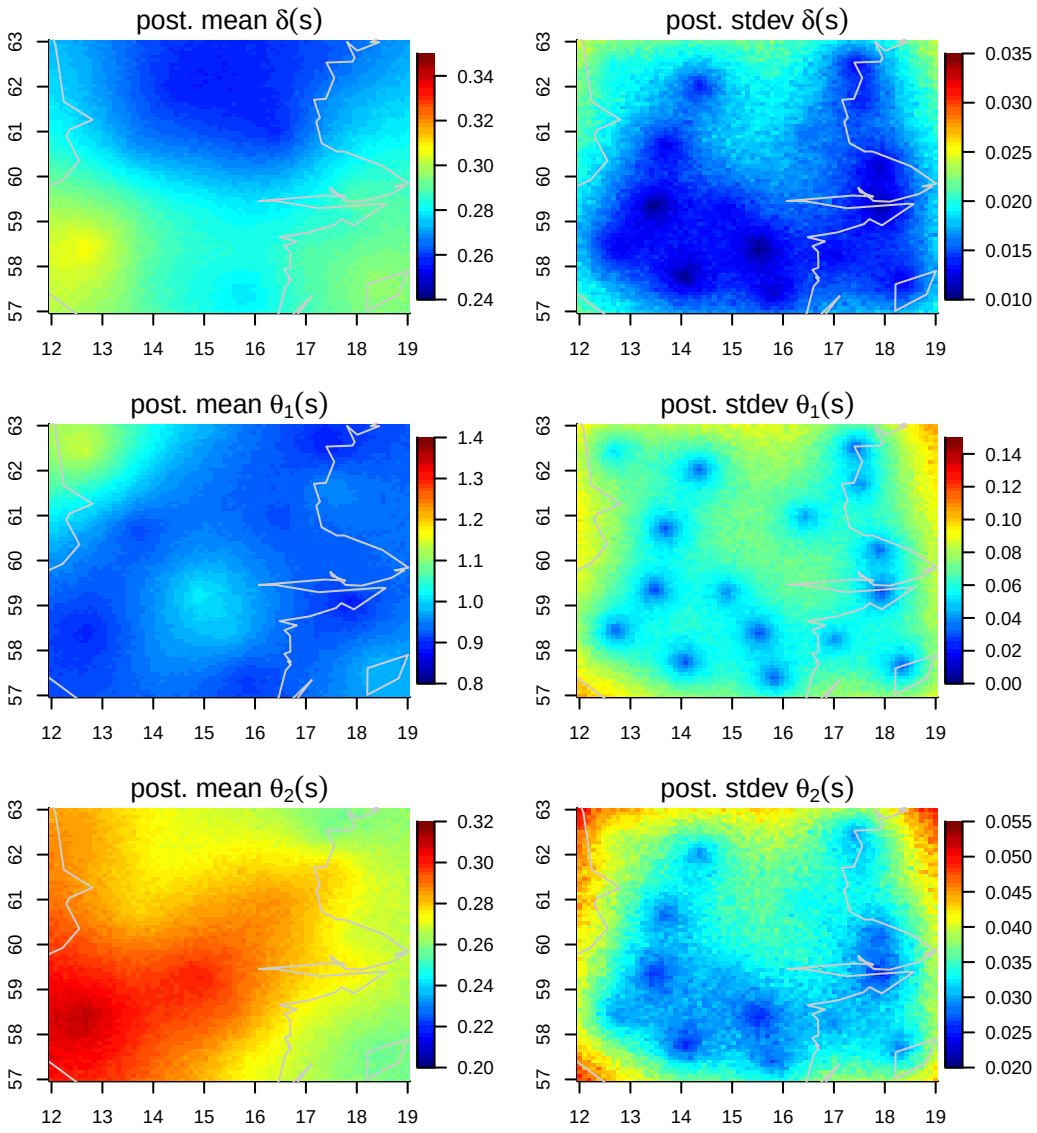

Figure 6: Posterior mean fields (left panels) and standard deviation fields (right panels) for the parameters characterizing the short and long range dependence in the noise.

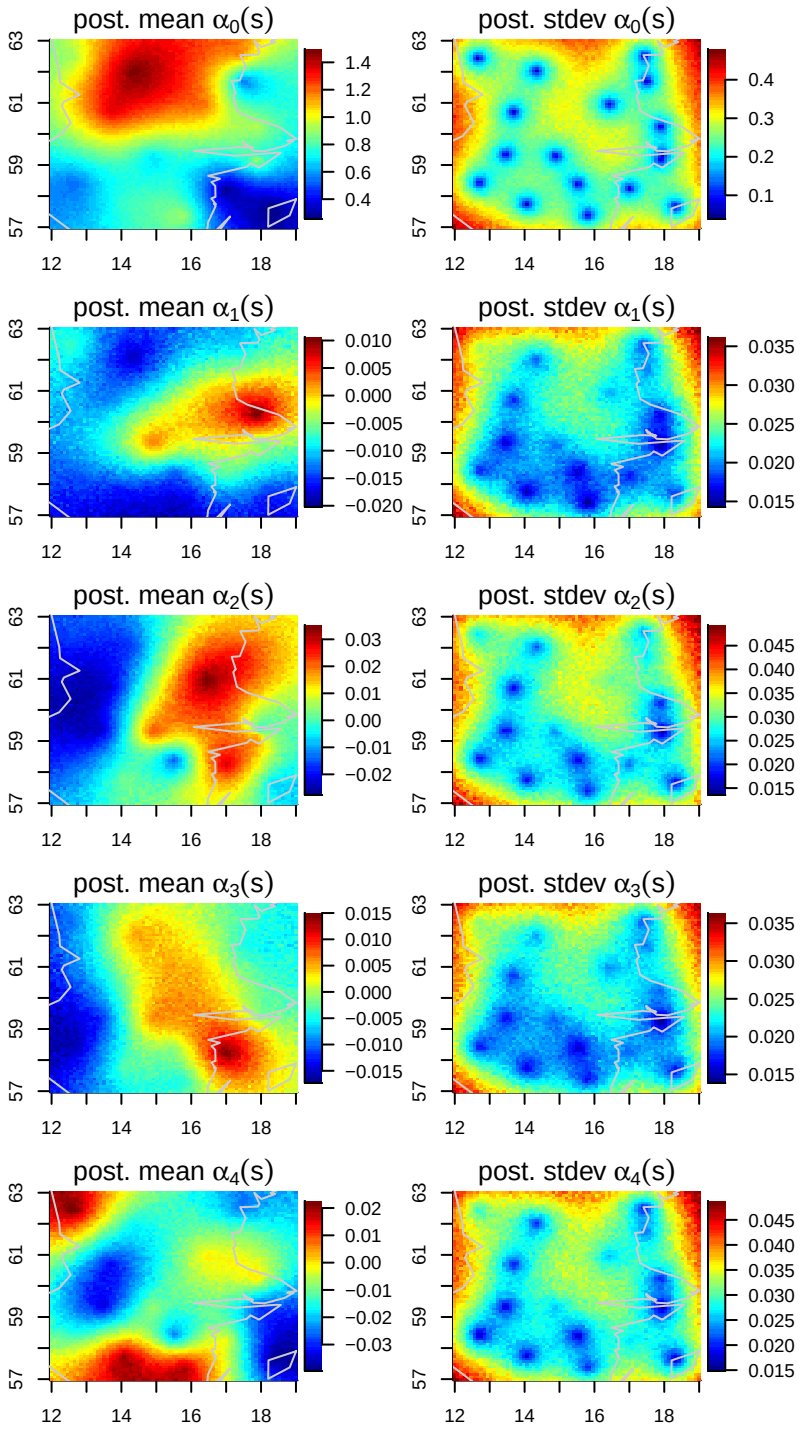

Figure 7: Posterior mean fields (left panels) and standard deviation fields (right panels) for the parameters $\{\alpha_j(\boldsymbol{s}) : j = 0, \ldots, 4\}$ characterizing the space-time volatility in daily temperature.

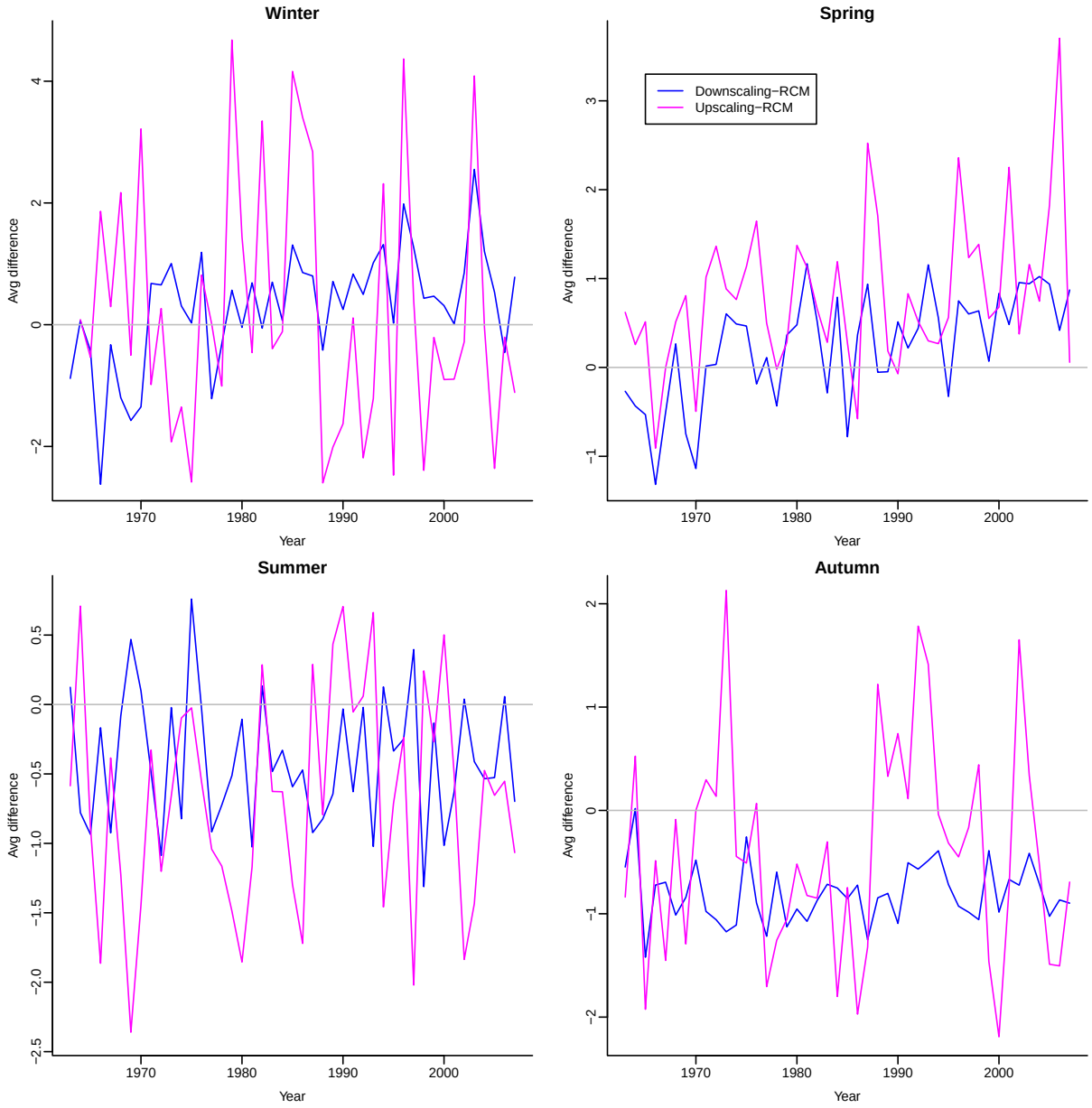

Figure 8: Average difference in the predicted seasonal fields as yielded by the downscaling and the RCM model (blue line) and the upscaling and the RCM model (magenta line) over time.