A. Metrics

This section describes the evaluation protocol and the set of metrics used to compare the performance of SerpentFlow against the baseline. Most of the metrics are inspired by the work produced in (François et al., 2020). The evaluation is performed separately against the reanalysis and the raw GCM (interpolated) output. All metrics are computed per climate variable and then averaged across the set of variables $\mathcal{V}$ considered in the study.

A.1. Notations

Let $X^{m,v}_{t,i,j}$ denote the value of climate variable $v \in \mathcal{V}$ produced by method m at time t and grid point (i, j). The following notations are used:

- $X^{\text{obs},v}_{t,i,j}$: reanalysis for variable v,
- $X^{\text{GCM},v}_{t,i,j}$: raw (interpolated) GCM output for variable v,
- $m \in \mathcal{M}$: downscaling method (CDF-t, R2D2, Dual FM, SerpentFlow variants),
- T : set of time steps,
- S : set of spatial grid points,
- $\mathcal{V}$: set of variables (e.g., wind speed, maximum wind speed, and the zonal and meridional wind components).

Spatial averages are taken over all grid points $(i, j) \in S$, and temporal averages over the evaluation period T. Metrics are then averaged across all climate variables $v \in \mathcal{V}$ unless otherwise specified.

A.2. Metrics relative to the observations

A.2.1. Mean and standard deviation differences

For each climate variable v, the mean difference between method m and observations is:

$$\Delta\mu^{m,v} = \frac{1}{|S|} \sum_{(i,j)\in S} \left|\mu^{m,v}_{i,j} - \mu^{\text{obs},v}_{i,j}\right|, \quad \mu^{m,v}_{i,j} = \frac{1}{|T|} \sum_{t\in T} X^{m,v}_{t,i,j} \tag{A.1}$$

The standard deviation difference is:

$$\Delta\sigma^{m,v} = \frac{1}{|S|} \sum_{(i,j)\in S} \left|\sigma^{m,v}_{i,j} - \sigma^{\text{obs},v}_{i,j}\right|, \quad \sigma^{m,v}_{i,j} = \sqrt{\frac{1}{|T|} \sum_{t\in T} \left(X^{m,v}_{t,i,j} - \mu^{m,v}_{i,j}\right)^2} \tag{A.2}$$

Finally, metrics are averaged across variables:

$$\Delta\mu^{m} = \frac{1}{|\mathcal{V}|} \sum_{v\in\mathcal{V}} \Delta\mu^{m,v}, \quad \Delta\sigma^{m} = \frac{1}{|\mathcal{V}|} \sum_{v\in\mathcal{V}} \Delta\sigma^{m,v}. \tag{A.3}$$

A.2.2. Inter-variable correlation consistency

For each downscaling method m, climate variable pair (v_1, v_2) and grid point (i, j), the temporal Pearson correlation is defined as

$$\rho^{m,v_1,v_2}_{i,j} = \frac{\sum_{t\in T} \left(X^{m,v_1}_{t,i,j} - \mu^{m,v_1}_{i,j}\right)\left(X^{m,v_2}_{t,i,j} - \mu^{m,v_2}_{i,j}\right)}{\sqrt{\sum_{t\in T} \left(X^{m,v_1}_{t,i,j} - \mu^{m,v_1}_{i,j}\right)^2 \sum_{t\in T} \left(X^{m,v_2}_{t,i,j} - \mu^{m,v_2}_{i,j}\right)^2}}, \tag{A.4}$$

where $\mu_{i,j}^{m,v} = \frac{1}{|T|}\sum_{t\in T} X_{t,i,j}^{m,v}$ is the temporal mean of variable v at grid point (i, j).

The mean absolute difference of Pearson correlations between method m and the observations is then computed as

$$\Delta\rho^{m,v_1,v_2} = \frac{1}{|S|}\sum_{(i,j)\in S}\left|\rho_{i,j}^{m,v_1,v_2} - \rho_{i,j}^{\mathrm{obs},v_1,v_2}\right|. \tag{A.5}$$

Finally, the metric is averaged over all distinct variable pairs to produce a single inter-variable correlation consistency score:

$$\Delta\rho^{m} = \frac{2}{|\mathcal{V}|(|\mathcal{V}|-1)}\sum_{v_1<v_2}\frac{1}{|S|}\sum_{(i,j)\in S}\left|\rho_{i,j}^{m,v_1,v_2} - \rho_{i,j}^{\mathrm{obs},v_1,v_2}\right| \tag{A.6}$$

A.2.3. Spatial Spearman correlation structure

For each variable v, we first define spatial anomalies by removing the spatial mean at each time step:

$$X_{t,i,j}^{\prime m,v} = X_{t,i,j}^{m,v} - \frac{1}{|S|}\sum_{(i,j)\in S} X_{t,i,j}^{m,v}. \tag{A.7}$$

Let $r_{t,i,j}^{m,v}$ denote the rank of $X_{t,i,j}^{\prime m,v}$ among all grid points (i, j) at fixed time t. The Spearman correlation matrix $\mathbf{R}^{m,v}$ is then defined by computing the Pearson correlation of the ranks across time for every pair of grid points (p, q):

$$\mathbf{R}_{p,q}^{m,v} = \frac{\sum_{t\in T}\left(r_{t,p}^{m,v} - \bar{r}_{p}^{m,v}\right)\left(r_{t,q}^{m,v} - \bar{r}_{q}^{m,v}\right)}{\sqrt{\sum_{t\in T}\left(r_{t,p}^{m,v} - \bar{r}_{p}^{m,v}\right)^2 \sum_{t\in T}\left(r_{t,q}^{m,v} - \bar{r}_{q}^{m,v}\right)^2}}, \quad \bar{r}_{p}^{m,v} = \frac{1}{|T|}\sum_{t\in T} r_{t,p}^{m,v}. \tag{A.8}$$

The spatial correlation error for variable v is then computed as the mean absolute difference with the observations:

$$\Delta R^{m,v} = \frac{1}{|S|^2}\sum_{p,q\in S}\left|\mathbf{R}_{p,q}^{m,v} - \mathbf{R}_{p,q}^{\mathrm{obs},v}\right|. \tag{A.9}$$

Finally, this metric is averaged across all variables to produce a single score:

$$\Delta\mathbf{R}^{m} = \frac{1}{|\mathcal{V}|}\sum_{v\in\mathcal{V}}\frac{1}{|S|^2}\sum_{p,q\in S}\left|\mathbf{R}_{p,q}^{m,v} - \mathbf{R}_{p,q}^{\mathrm{obs},v}\right| \tag{A.10}$$

A.2.4. Spatial power spectrum metric

For each climate variable v, the isotropic spatial power spectrum is computed over the grid, and averaged over time. A scalar metric quantifies the relative large-scale energy difference between the downscaled field and observations:

$$\mathrm{SSM}^{m,v} = \frac{\sqrt{\frac{1}{|k\le k_{\max}|}\sum_{k\le k_{\max}}(P_k^{m,v} - P_k^{\mathrm{obs},v})^2}}{\frac{1}{|k\le k_{\max}|}\sum_{k\le k_{\max}} P_k^{\mathrm{obs},v}}, \quad \mathrm{SSM}^{m} = \frac{1}{|\mathcal{V}|}\sum_{v\in\mathcal{V}}\mathrm{SSM}^{m,v} \tag{A.11}$$

where $P_k^{m,v}$ denotes the mean power spectrum in wavenumber bin k, and $k_{\max}$ is a cutoff separating large-scale features.

A.2.5. Spatial variogram metric (mountains)

For mountainous regions (elevation > θ_{alt}), the semi-variogram $\gamma(h)$ is computed for each variable v, and compared to the observations:

$$\gamma^{m,v}(h) = \frac{1}{2|(i,j) \in \text{bin } h|} \sum_{(i,j)\in \text{bin } h} \left(X_i^{m,v} - X_j^{m,v}\right)^2, \tag{A.12}$$

$$\text{OVM}^m = \frac{1}{|\mathcal{V}|} \sum_{v\in\mathcal{V}} \frac{\sqrt{\frac{1}{h\leq h_{\max}} \sum_{h\leq h_{\max}} (\gamma^{m,v}(h) - \gamma^{\text{obs},v}(h))^2}}{\frac{1}{h\leq h_{\max}} \sum_{h\leq h_{\max}} \gamma^{\text{obs},v}(h)} \tag{A.13}$$

where h denotes distance bins and $h_{\max}$ a cutoff distance.

A.2.6. Kolmogorov–Smirnov statistic

For each variable v, let $X_{t,i,j}^{m,v}$ denote the values produced by method m at all times $t \in T$ and all spatial points $(i, j) \in S$. The empirical cumulative distribution function (CDF) of method m is defined as

$$F^{m,v}(x) = \frac{1}{|T||S|} \sum_{t\in T} \sum_{(i,j)\in S} \mathbb{I}\left(X_{t,i,j}^{m,v} \leq x\right), \tag{A.14}$$

where $\mathbb{I}$ is the indicator function. Similarly, $F^{\text{obs},v}(x)$ is the empirical CDF of the observations' values for variable v.

The two-sample Kolmogorov–Smirnov (KS) statistic between method m and the observations for variable v is then defined as

$$\text{KS}^{m,v} = \sup_x \left|F^{m,v}(x) - F^{\text{obs},v}(x)\right|. \tag{A.15}$$

Finally, the KS statistic is averaged over all variables:

$$\text{KS}^m = \frac{1}{|\mathcal{V}|} \sum_{v\in\mathcal{V}} \sup_x \left|F^{m,v}(x) - F^{\text{obs},v}(x)\right| \tag{A.16}$$

A.2.7. Extreme value metrics

Using variable-specific observations thresholds $\theta_{i,j}^v = Q_{0.95}(X_{\cdot,i,j}^{\text{obs},v})$:

Extreme frequency.

$$f_{i,j}^{m,v} = \frac{1}{|T|} \sum_{t\in T} \mathbb{I}(X_{t,i,j}^{m,v} > \theta_{i,j}^v), \quad f^m = \frac{1}{|\mathcal{V}|} \sum_{v\in\mathcal{V}} \frac{1}{|S|} \sum_{(i,j)\in S} |f_{i,j}^{m,v} - f_{i,j}^{\text{obs},v}| \tag{A.17}$$

Extreme intensity.

$$I_{i,j}^{m,v} = \frac{1}{N_{i,j}^{m,v}} \sum_{t:X_{t,i,j}^{m,v} > \theta_{i,j}^v} X_{t,i,j}^{m,v}, \quad I^m = \frac{1}{|\mathcal{V}|} \sum_{v\in\mathcal{V}} \frac{1}{|S|} \sum_{(i,j)\in S} |I_{i,j}^{m,v} - I_{i,j}^{\text{obs},v}| \tag{A.18}$$

A.3. Metrics relative to the GCM

A.3.1. Temporal correlation

For each variable v, the temporal Pearson correlation with the GCM is:

$$\rho_{\text{GCM},i,j}^{m,v} = \text{corr}_t(X_{\cdot,i,j}^{m,v}, X_{\cdot,i,j}^{\text{GCM},v}), \quad \rho_{\text{GCM}}^m = \frac{1}{|\mathcal{V}|} \sum_{v\in\mathcal{V}} \frac{1}{|S|} \sum_{(i,j)\in S} \text{corr}_t(X_{\cdot,i,j}^{m,v}, X_{\cdot,i,j}^{\text{GCM},v}) \tag{A.19}$$

A.3.2. Global annual anomalies

Let $\overline{X}_y^{m,v} = \frac{1}{|S|}\sum_{(i,j)\in S} X_{\text{year }y,i,j}^{m,v}$. Annual anomaly deviations:

$$A_y^{m,v} = \overline{X}_y^{m,v} - \frac{1}{N_y}\sum_y \overline{X}_y^{m,v}, \quad A^m = \frac{1}{|\mathcal{V}|}\sum_{v\in\mathcal{V}} \frac{1}{N_y}\sum_y |A_y^{m,v} - A_y^{\text{GCM},v}| \tag{A.20}$$

A.3.3. Relative climate change signal

For each variable v, the relative change between a historical period and a future period is computed as

$$\Delta_{\text{full}}^{m,v} = \frac{\overline{X}_{\text{fut}}^{m,v} - \overline{X}_{\text{hist}}^{m,v}}{\overline{X}_{\text{hist}}^{m,v}} \times 100, \tag{A.21}$$

where $\overline{X}_{\text{hist}}^{m,v}$ and $\overline{X}_{\text{fut}}^{m,v}$ are spatial averages over the historical and future periods, respectively.

For the seasonal variant, the same metric is computed for each meteorological season s:

- DJF: December–January–February,
- MAM: March–April–May,
- JJA: June–July–August,
- SON: September–October–November.

The seasonal relative change is then

$$\Delta_{\text{season},s}^{m,v} = \frac{\overline{X}_{\text{fut},s}^{m,v} - \overline{X}_{\text{hist},s}^{m,v}}{\overline{X}_{\text{hist},s}^{m,v}} \times 100, \tag{A.22}$$

with averages computed over the months corresponding to season s.

Finally, metrics are averaged over all variables and, for the seasonal metric, over all seasons:

$$\Delta_{\text{full}}^{m} = \frac{1}{|\mathcal{V}|}\sum_{v\in\mathcal{V}} |\Delta_{\text{full}}^{m,v} - \Delta_{\text{full}}^{\text{GCM},v}|, \tag{A.23}$$

$$\Delta_{\text{season}}^{m} = \frac{1}{|\mathcal{V}|}\sum_{v\in\mathcal{V}} \frac{1}{|S|}\sum_{(i,j)\in S} \frac{1}{4}\sum_{s\in\{\text{DJF,MAM,JJA,SON}\}} |\Delta_{\text{season},s}^{m,v} - \Delta_{\text{season},s}^{\text{GCM},v}|. \tag{A.24}$$

A.4. Summary

All metrics are computed per variable, averaged over space, and then aggregated across variables to produce method-level scores.

Below are the values for each metric for the two experiments detailed Section 3.

Table 1. *Summary of evaluation metrics*

Metric	Description / Equation
Metrics relative to the observations	
$\Delta\mu$	Mean difference (Eq. A.1)
$\Delta\sigma$	Standard deviation difference (Eq. A.2)
$\Delta\rho$	Inter-variable temporal correlation consistency (Eq. A.6)
ΔR	Spatial Spearman correlation error (Eq. A.10)
SSM	Spatial power spectrum metric (large-scale energy)
OVM	Orographic variogram metric (spatial variability in mountains)
KS	Kolmogorov–Smirnov statistic (CDF error, Eq. A.16)
f	Extreme frequency (Eq. A.17)
I	Extreme intensity (Eq. A.18)
Metrics relative to GCM	
ρ_{GCM}	Temporal correlation with GCM (Eq. A.19)
A	Annual anomaly deviations (inter-annual variability, Eq. A.20)
Δ_{full}	Full-period relative climate change signal (Eq. A.21)
Δ_{season}	Seasonal relative climate change signal (Eq. A.24)

Table 2. *Performance metrics with respect to ERA5 observations. For ensemble methods, values are reported as mean ± standard deviation across members.*

Metric	CDF-t	R2D2	DFM	DFMm	SF12	SF12m	SF7	SF3	GCM
Distance to CDF	0.016	0.016	0.024	0.024	0.018	**0.011**	0.013	0.012	0.062
Spatial mean	0.109	0.098	$0.064_{\pm 0.000}$	0.064	$\mathbf{0.060_{\pm 0.001}}$	0.061	$0.061_{\pm 0.000}$	$0.063_{\pm 0.000}$	0.639
Spatial standard deviation	0.062	0.074	0.068	0.068	$0.089_{\pm 0.000}$	0.057	$0.056_{\pm 0.000}$	$\mathbf{0.055_{\pm 0.000}}$	0.452
Inter variables correlation	0.121	**0.024**	0.108	0.108	$0.069_{\pm 0.000}$	0.056	$0.091_{\pm 0.000}$	$0.103_{\pm 0.000}$	0.128
Spatial spearman correlation	19350	42407	$14504_{\pm 0}$	14504	$14931_{\pm 33}$	**10055**	$15733_{\pm 20}$	$17513_{\pm 7}$	30929
Spatial Variogram (mountains)	0.227	0.705	0.067	0.067	$0.129_{\pm 0.000}$	**0.033**	$0.131_{\pm 0.000}$	$0.162_{\pm 0.000}$	0.305
Spatial spectrum	0.188	0.540	$\mathbf{0.078_{\pm 0.000}}$	0.078	$0.173_{\pm 0.001}$	0.089	$0.203_{\pm 0.001}$	$0.215_{\pm 0.000}$	0.719
Extreme frequency	0.006	0.007	0.011	0.011	$0.007_{\pm 0.000}$	**0.006**	$0.007_{\pm 0.000}$	$0.007_{\pm 0.000}$	0.052
Extreme intensity	**0.103**	0.118	0.200	0.200	$0.116_{\pm 0.001}$	0.118	$0.137_{\pm 0.000}$	$0.158_{\pm 0.001}$	0.349

DFM = Dual FM, DFMm = Dual FM mbr, SF12 = SF 1200 km, SF12m = SF 1200 km mbr, SF7 = SF 750 km, SF3 = SF 300 km

Table 3. *Performance metrics with respect to GCM. For ensemble methods, values are reported as mean ± standard deviation across members.*

Metric	CDF-t	R2D2	DFM	DFMm	SF12	SF12m	SF7	SF3	GCM
Temporal correlation	**0.996**	0.473	0.715	0.715	$0.893_{\pm 0.000}$	0.853	$0.932_{\pm 0.000}$	$0.956_{\pm 0.000}$	–
Inter-annual variability	**0.012**	0.088	$0.066_{\pm 0.000}$	0.066	$0.013_{\pm 0.000}$	0.013	$0.014_{\pm 0.000}$	$0.017_{\pm 0.000}$	–
Delta mean (annual)	**0.167**	0.684	$0.880_{\pm 0.000}$	0.880	$0.625_{\pm 0.012}$	0.636	$0.620_{\pm 0.008}$	$0.514_{\pm 0.006}$	–
Delta mean (seasonal)	**0.266**	1.097	1.516	1.516	$1.012_{\pm 0.018}$	1.098	$0.921_{\pm 0.010}$	$0.757_{\pm 0.006}$	–

DFM = Dual FM, DFMm = Dual FM mbr, SF12 = SF 1200 km, SF12m = SF 1200 km mbr, SF7 = SF 750 km, SF3 = SF 300 km

Table 4. *Performance metrics with respect to ERA5 observations (sfcWind variable only). For ensemble methods, values are reported as mean ± standard deviation across members.*

Metric	CDF-t	R2D2	DFM	DFMm	SF12	SF12m	SF7	SF3	GCM
Distance to CDF	0.014	0.014	0.032	0.032	0.013	**0.005**	0.010	0.009	0.061
Spatial Variogram (mountains)	0.170	0.190	0.026	0.026	$0.083_{\pm 0.000}$	**0.024**	$0.084_{\pm 0.000}$	$0.111_{\pm 0.000}$	0.206
Spatial spectrum	0.124	0.139	0.054	0.054	$0.090_{\pm 0.003}$	**0.038**	$0.090_{\pm 0.001}$	$0.104_{\pm 0.000}$	0.761
Extreme frequency	0.006	0.006	0.008	0.008	**0.005**	0.005	0.005	0.005	0.061
Extreme intensity	0.103	0.103	0.172	0.172	**0.087**	0.092	0.089	0.102	0.333

DFM = Dual FM, DFMm = Dual FM mbr, SF12 = SF 1200 km, SF12m = SF 1200 km mbr, SF7 = SF 750 km, SF3 = SF 300 km

B. Ensemble calibration

As described in Section 2, SerpentFlow generates downscaled fields by sampling stochastic noise in the domain-specific (high-frequency) component and passing it through the learned generator f_θ. At inference, this noise is drawn as $z \sim \mathcal{N}(\mathbf{0}, a^2\mathbf{I})$, where the scaling factor a controls the amplitude of the injected variability. When $a = 1$, the inference noise matches the unit-variance distribution seen during training. Adjusting a allows post-hoc control over ensemble spread without retraining: lower values produce tighter ensembles, while higher values increase diversity among members.

We generate $N = 10$ ensemble members for each time step on ERA5 low-passed values (perfect model setup) and evaluate calibration against the ground truth over the 2000–2020 period using four standard diagnostics: the mean Continuous Ranked Probability Score (CRPS), the spread–skill ratio, rank histograms, and reliability diagrams. All diagnostics are shown for the four wind channels (`sfcWind`, `uas`, `vas`, `sfcWindmax`), with the rank histogram and reliability diagram displayed for `sfcWind`.

***a* = 1.0** *(Figure 4).*. The rank histogram exhibits a clear U-shape, indicating that the ensemble is underdispersive: ERA5 values frequently fall outside the ensemble range. Consistently, the ensemble spread is noticeably smaller than the RMSE across all channels, and the reliability curve lies above the diagonal at high nominal levels, confirming that prediction intervals are too narrow. This suggests that

Table 5. *Performance metrics with respect to GCM (sfcWind variable only). For ensemble methods, values are reported as mean ± standard deviation across members.*

Metric	CDF-t	R2D2	DFM	DFMm	SF12	SF12m	SF7	SF3	GCM
Temporal correlation	**0.997**	0.835	0.646	0.646	0.879	0.837	0.922	0.949	–
Inter-annual variability	0.006	0.012	0.077	0.077	0.005	**0.005**	0.007	0.010	–
Delta mean (annual)	**0.138**	0.511	0.895	0.895	0.608	0.583	0.543	0.443	–
Delta mean (seasonal)	**0.175**	0.802	1.057	1.057	1.037	1.085	0.712	0.608	–

DFM = Dual FM, DFMm = Dual FM mbr, SF12 = SF 1200 km, SF12m = SF 1200 km mbr, SF7 = SF 750 km, SF3 = SF 300 km

Table 6. *Performance metrics with respect to SAFRAN*

Metric	CDF-t	RCM adjusted	SF 750 km	SF 600 km	SF 500 km	SF 300 km	GCM
Distance to CDF	**0.032**	0.042	0.035	0.034	0.036	0.036	0.212
Spatial mean	0.320	**0.196**	0.388	0.389	0.388	0.391	1.147
Spatial standard deviation	0.180	**0.099**	0.235	0.236	0.234	0.235	0.532
Spatial spearman correlation	0.166	**0.091**	0.111	0.114	0.108	0.131	0.245
Extreme frequency	0.014	**0.008**	0.020	0.020	0.020	0.021	0.070
Extreme intensity	0.223	**0.128**	0.237	0.237	0.249	0.267	0.603

Table 7. *Performance metrics with respect to GCM (SAFRAN experimentation)*

Metric	CDF-t	RCM adjusted	SF 750 km	SF 600 km	SF 500 km	SF 300 km
Temporal correlation	**0.957**	0.085	0.818	0.811	0.853	0.920
Inter-annual variability	**0.032**	0.200	0.058	0.064	0.050	0.047
Delta mean (annual)	**0.423**	2.077	0.711	0.884	0.569	1.209
Delta mean (seasonal)	1.068	3.544	1.122	1.316	**0.955**	1.537

the unit-variance noise used during training does not fully capture the variability needed at inference, likely because the generator learns a partial mapping from noise to fine-scale structure, reducing the effective stochasticity of the outputs.

***a* = 1.1** *(Figure 5)..* A moderate increase in noise amplitude substantially improves calibration. The rank histogram becomes nearly flat, the spread–skill gap narrows across all variables, and the reliability curve closely follows the diagonal. This configuration yields the best overall calibration among the three tested values, indicating that a slight amplification of the domain-specific noise compensates for the variance reduction introduced by the deterministic component of the generator.

***a* = 1.2** *(Figure 6)..* Further increasing the noise amplitude leads to overdispersion. The rank histogram develops a hump shape with depleted tails, indicating that ensemble members spread too widely. The spread now approaches or exceeds the RMSE for several channels, and the reliability curve falls below the diagonal, meaning that prediction intervals are wider than necessary. At this level, the injected noise dominates over the learned fine-scale structure, degrading the physical plausibility of individual members.

Based on this analysis, the value $a = 1.1$ should be used as the default noise scaling factor. This simple post-hoc tuning of the noise amplitude provides an effective and computationally inexpensive way to calibrate probabilistic predictions from SerpentFlow without any retraining.

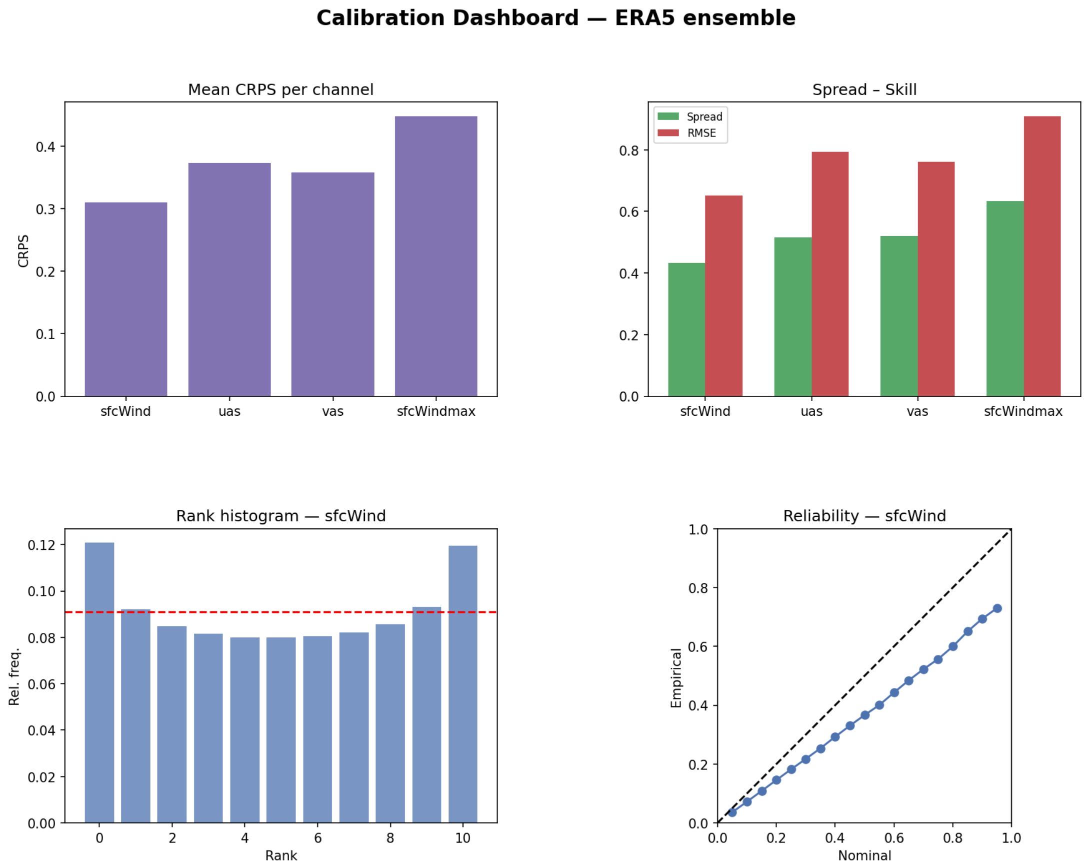


Figure 4. *Calibration dashboard for the ERA5 ensemble with noise scaling $a = 1.0$. The U-shaped rank histogram and the spread–skill gap indicate underdispersion..*

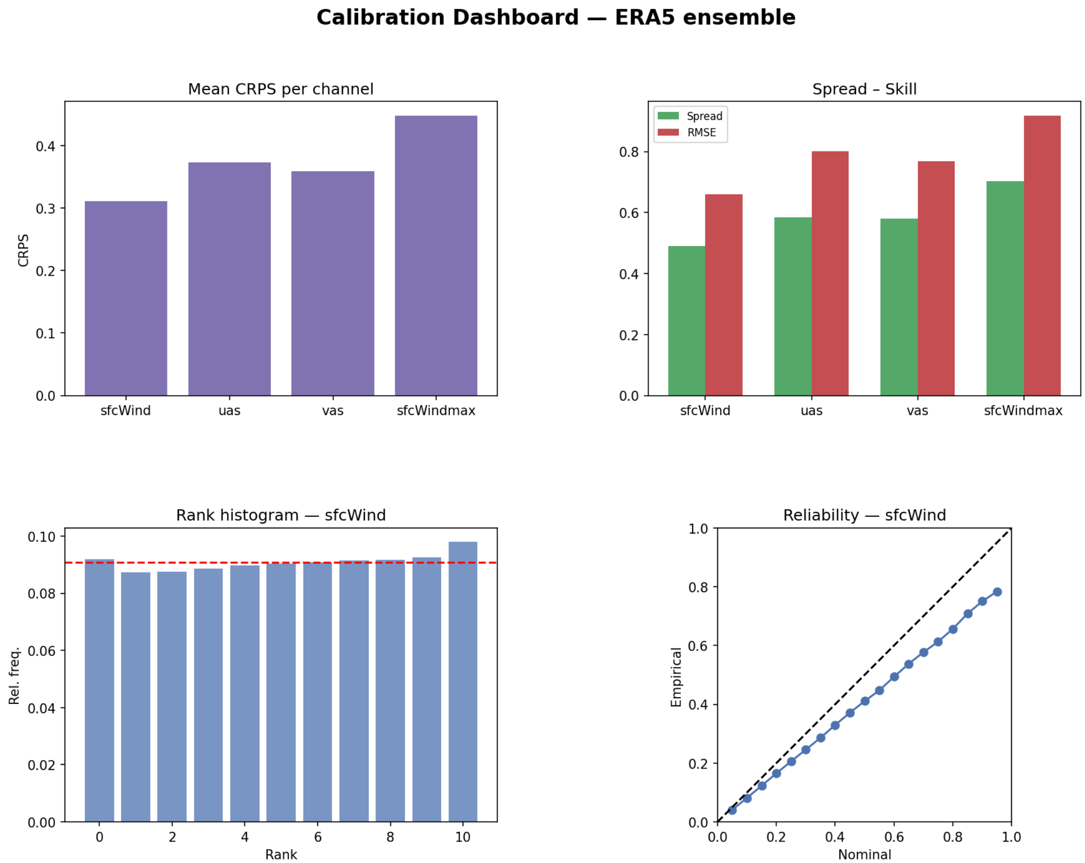


Figure 5. *Calibration dashboard for $a = 1.1$. The near-flat rank histogram, balanced spread–skill ratio, and diagonal reliability curve indicate well-calibrated ensembles..*

C. Additional Plots ACCESS/ERA5

C.1. Plots vs observations

C.2. Plots vs GCM

C.3. Deep learning based methods plots

D. Additional Plots CNRM/SAFRAN

D.1. Plots vs observations

D.2. Plots vs GCM

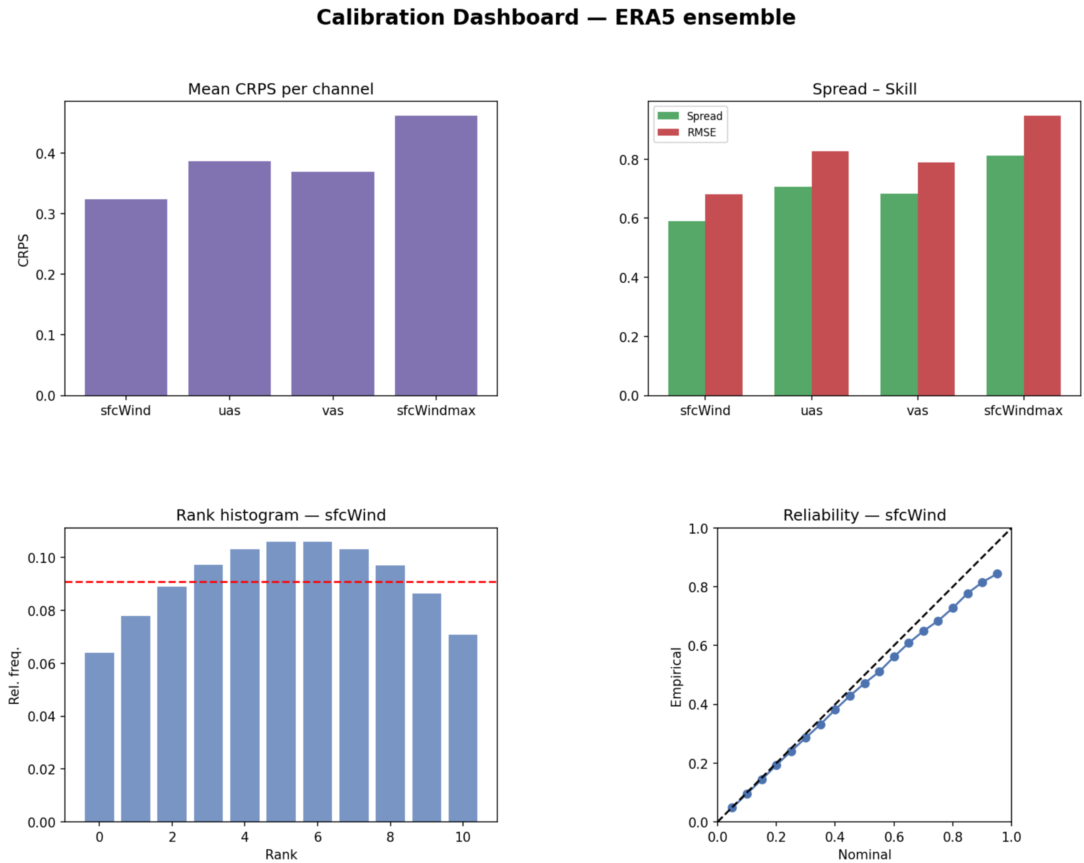
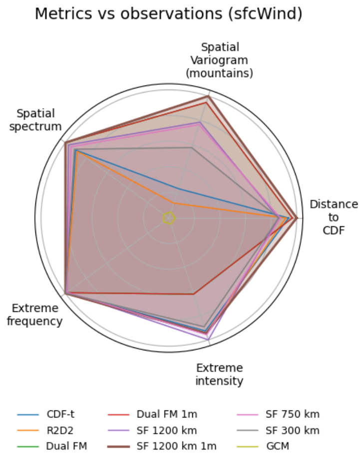
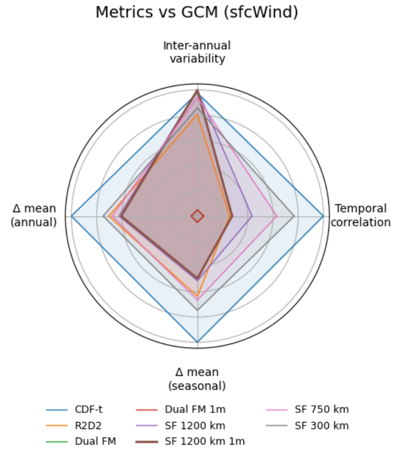


Figure 6. Calibration dashboard for $a = 1.2$. The hump-shaped rank histogram and the reliability curve below the diagonal indicate overdispersion..



(a) Metrics evaluated against ERA5. Results are detailed Table 4.



(b) Metrics evaluated against the GCM. Results are detailed Table 5.

Figure 7. *Radar plots summarizing the performance of all methods on the mean wind speed only. Higher values (closer to the outer circle) indicate better agreement with the reference. SF stands for SerpentFlow in all the plots, and "mbr" indicates "one member" for a generative method, the average of the members being shown otherwise.*

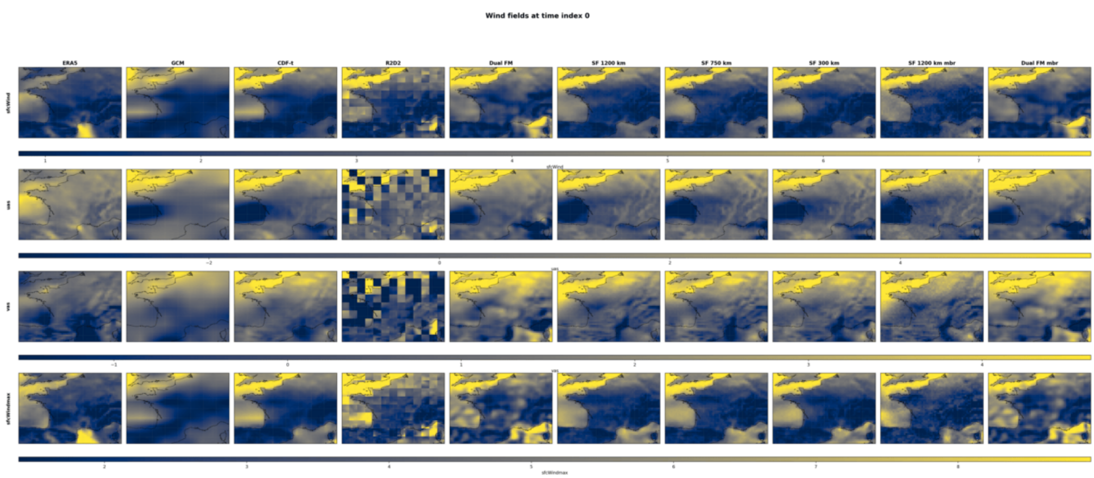
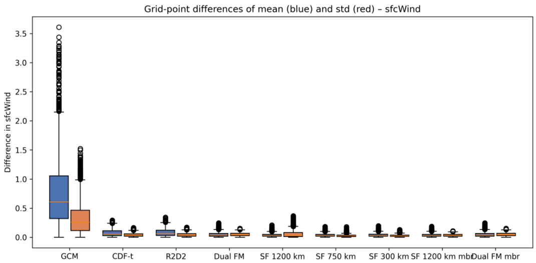
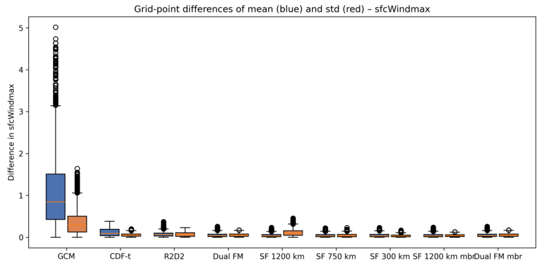


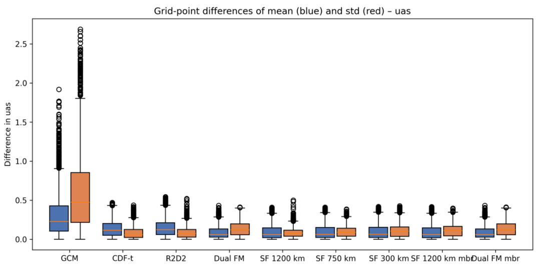
Figure 8. *Wind maps for the first time step, for each climate variables and for each method. The data and methods in the columns are, respectively: ERA5, GCM, CDF-t, R2D2, Dual FM, SerpentFlow (SF) 1200 km, SF 750 km, SF 300 km, SF 1200 km (1 member only), and Dual FM (1 member only). The rows show sfcWind, uas, vas, and sfcWindmax, respectively. The squares appearing on the R2D2 plot for each climate variable are due to the 5×5 spatial patches that had to be used due to limitations of the method. ERA5 and GCM are not aligned in time, which explains the differences between the first two columns. For the different downscaling methods, we can see that the pattern given by the GCM is generally followed. Dual FM seems to have created some artifacts, particularly in Brittany (in the west).*



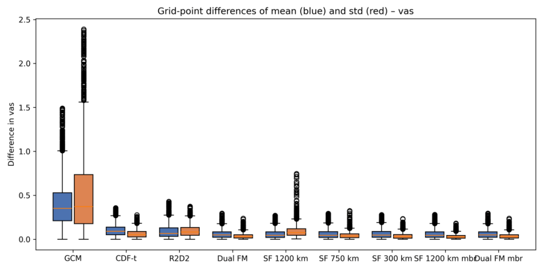
(a) Wind Speed



(b) Maximum wind speed

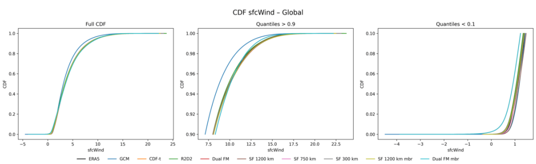


(c) Zonal wind

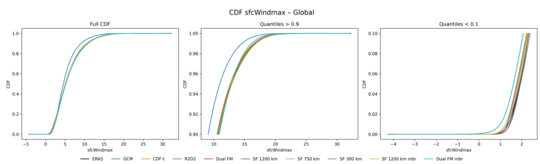


(d) Meridional wind

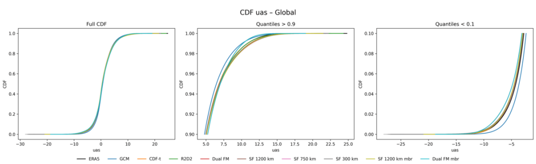
Figure 9. *Distribution of mean and standard deviations differences w.r.t ERA5 per grid point. All methods achieved spatially consistent bias correction, obtaining deviations from means and standard deviations well below those of the GCM. This is particularly true for wind speed variables. It is also noted that R2D2 deviates quite significantly from the mean for the meridional wind, which may explain the performance shown on the radar plot. Similarly, SF 1200 km (the mean) shows some deviations for the standard deviation of the meridional wind.*



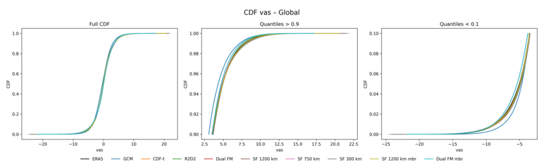
(a) Wind Speed



(b) Maximum wind speed

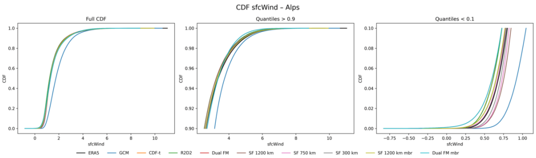


(c) Zonal wind

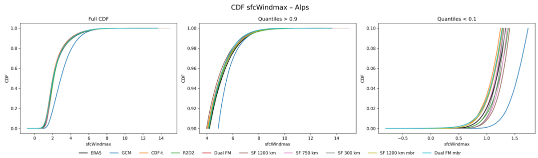


(d) Meridional wind

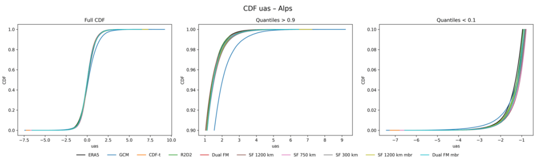
Figure 10. *Cumulative Distribution Functions (CDFs) for each method over the validation period. For each sublot a zoom is done on the extreme quantiles ($q \leq 0.1$ and $q \geq 0.9$). The bias correction is also clearly visible on these CDF plots. While the GCM tends to slightly underestimate wind speeds, downscaling methods correct this bias even in extreme quantiles. As in the paper (Keisler et al., 2026), we note that Dual FM produces negative values. The correction of u and v is more complicated in the high and low quantiles. Dual FM in particular fails.*



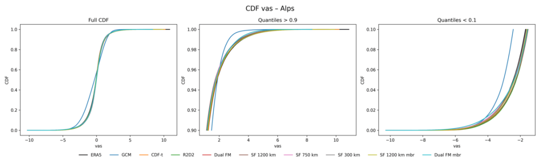
(a) Wind Speed



(b) Maximum wind speed

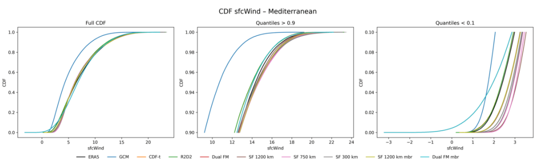


(c) Zonal wind

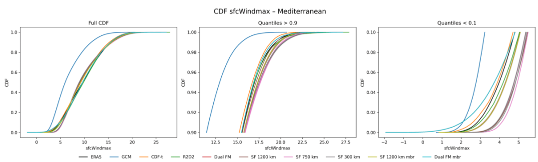


(d) Meridional wind

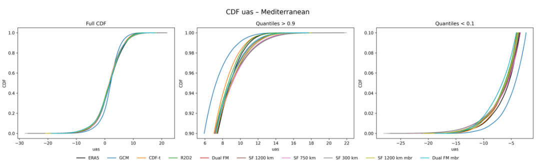
Figure 11. *Cumulative distribution functions (CDFs) for each method over a small region in the Alps (*latitude $\in [44, 47]$, longitude $\in [5, 8]$*). For each subplot, a zoom is applied to the extreme quantiles ($q \leq 0.1$ and $q \geq 0.9$). In this small region, the corrections remain very good, particularly for u and v. All methods are therefore successful in modeling marginal distributions. However, for wind speed variables, we note that SF 1200 km (single-member version) is significantly more accurate for the extreme quantiles (high and low) of ERA5.*



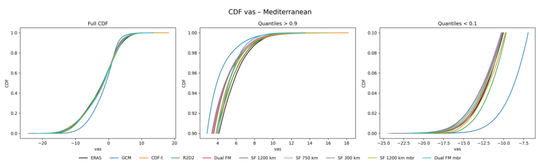
(a) Wind Speed



(b) Maximum wind speed



(c) Zonal wind



(d) Meridional wind

Figure 12. *Cumulative distribution functions (CDFs) for each method over a small region in the Mediterranean (*latitude $\in [41, 43]$, longitude $\in [3, 6]$*). For each subplot, a zoom is applied to the extreme quantiles ($q \leq 0.1$ and $q \geq 0.9$). In this region, R2D2 performs best in the high quantiles of all variables, while Dual FM performs best in the low quantiles. Nevertheless, in general, the distributions have been well corrected compared to the GCM.*

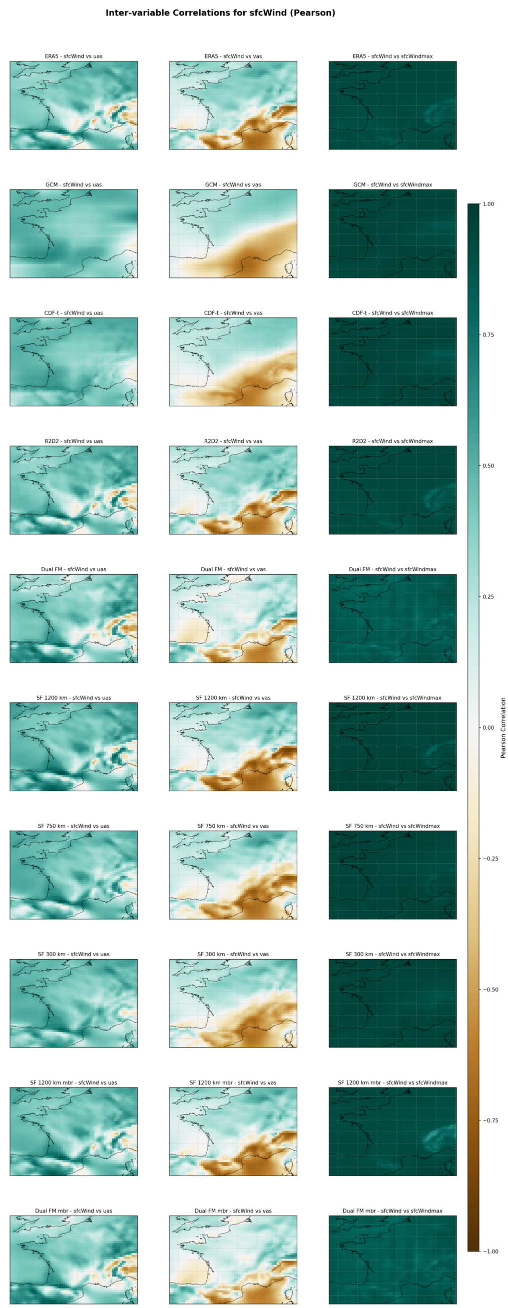


Figure 13. *Pearson correlation coefficients between wind speed and all other climate variables, by grid point and method. The columns show correlations with zonal wind, meridional wind, and maximum wind speed, respectively. The CDF-t correlations are very close to those of the GCM. The maps lean heavily toward green, indicating strong correlation between variables. However, for wind speed versus meridional wind in the ERA5 data, relief features and coastlines tend toward anti-correlation (dark brown). We also verified the physical consistency constraint sfcWind $\leq$ sfcWindmax: this holds for nearly all grid points and time steps, with violation rates comparable to those found in the raw GCM output..*

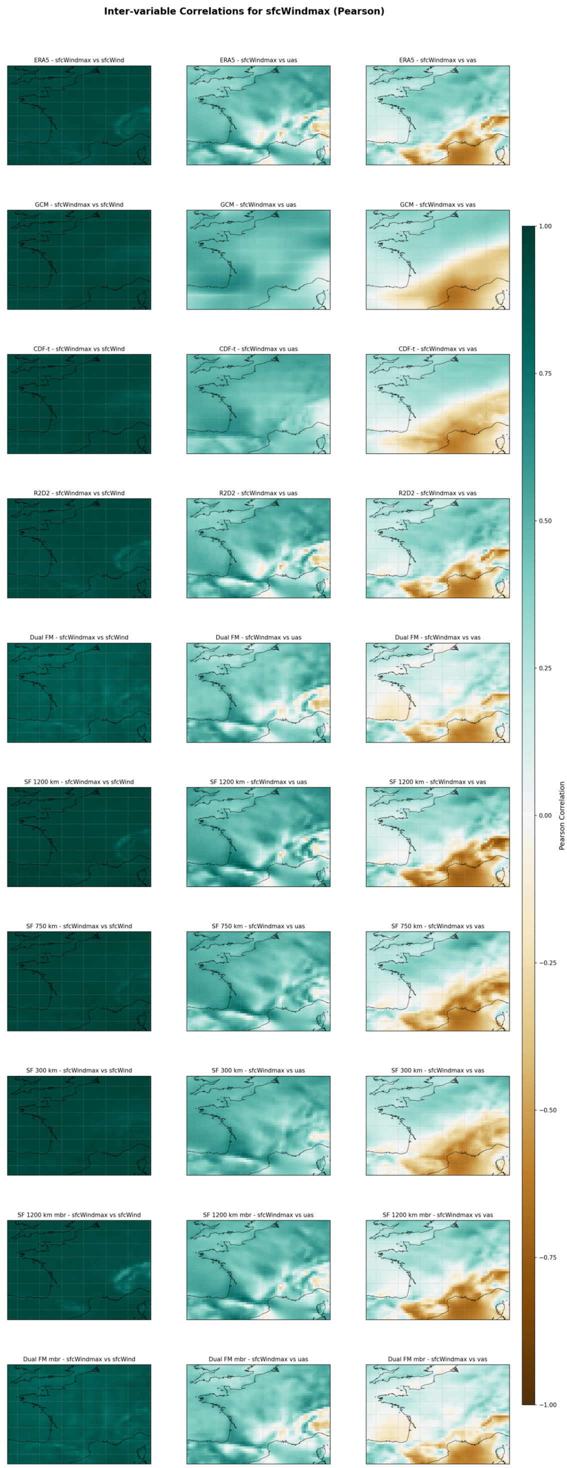


Figure 14. *Pearson correlations coefficient between maximum wind speed and all the other climate variables, by grid point and method. Respectively, the columns are correlations vs wind speed, zonal and meridional wind. Dual FM appears to be significantly less effective than SerpentFlow or R2D2 at correctly correcting inter-variable correlations. Between average wind speed and zonal wind, excessive correlations appear in flat areas. Between average and maximum speed, Dual FM tends to accentuate the loss of correlation over relief areas.*

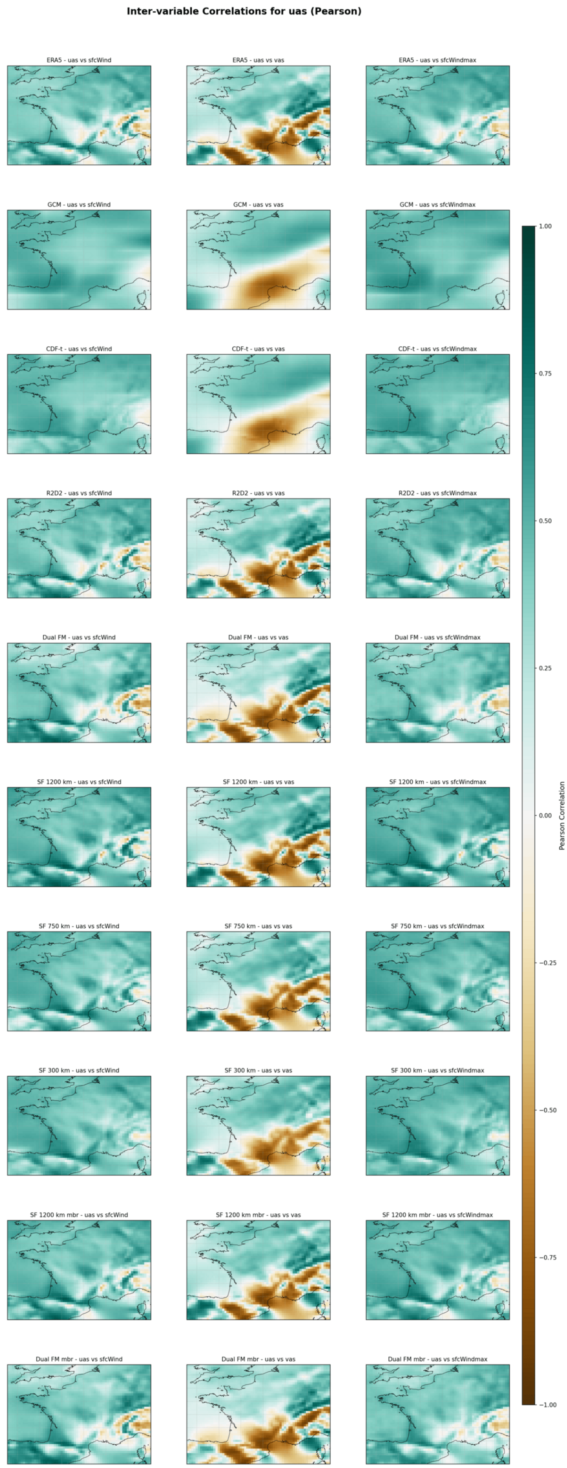


Figure 15. *Pearson correlations coefficient between zonal wind and all the other climate variables, by grid point and method. Respectively, the columns are correlations vs wind speed, meridional wind and max wind. This figure clearly shows, particularly in the middle column (u vs. v), that the lower the SerpentFlow cutoff frequency (i.e., the more small-scale elements are removed), the closer the correlation with ERA5. SF 300 km is much less accurate here than SF 1200 km.*

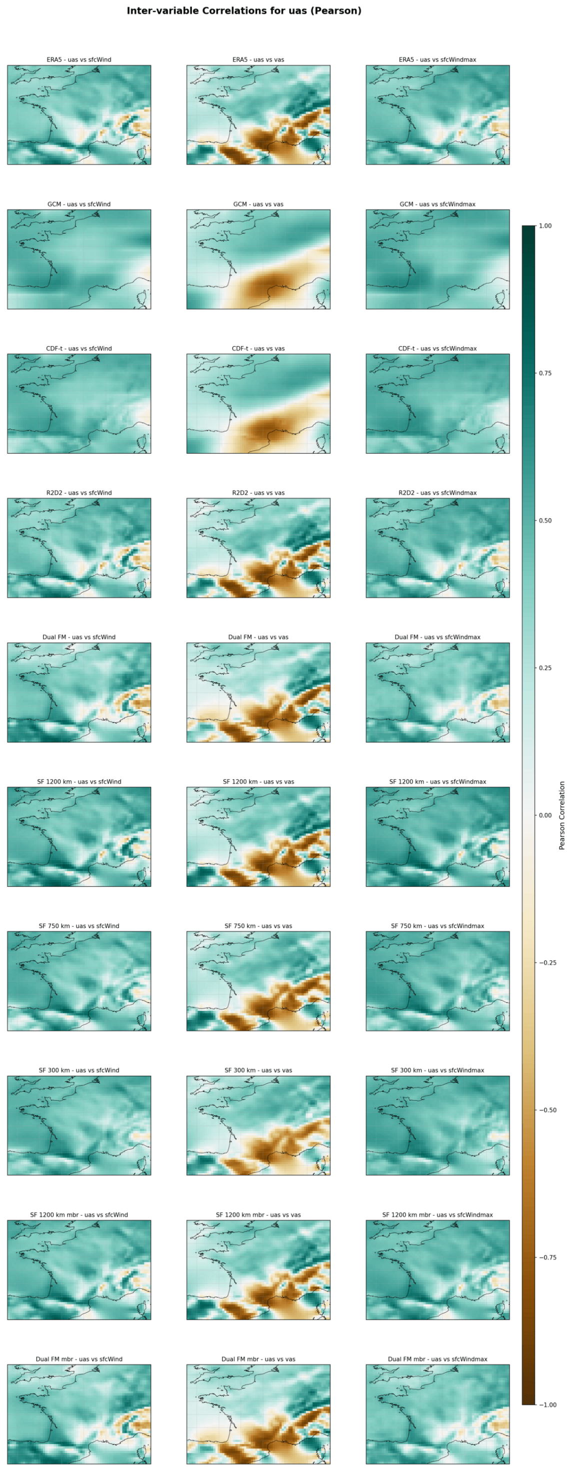
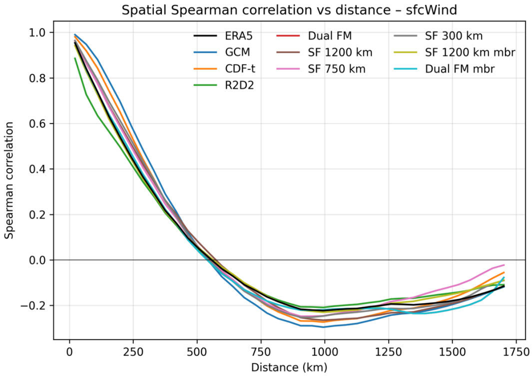
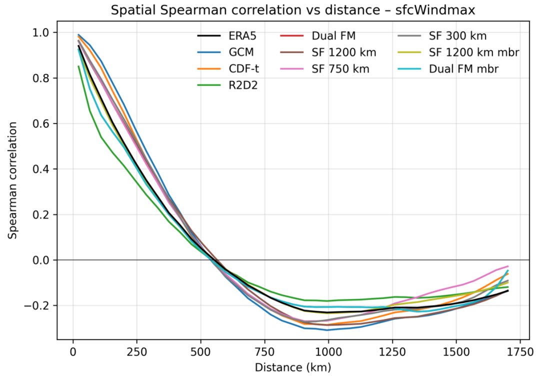


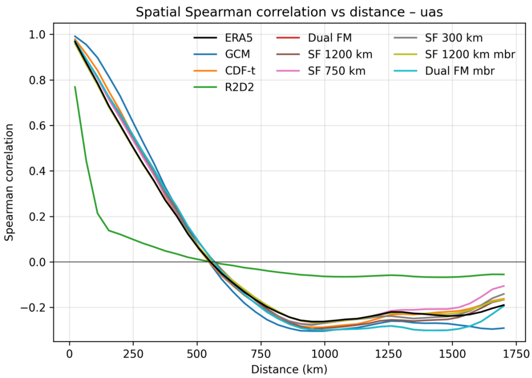
Figure 16. *Pearson correlations coefficient between meridional wind and all the other climate variables, by grid point and method. Respectively, the columns are correlations vs wind speed, zonal wind and max wind.*



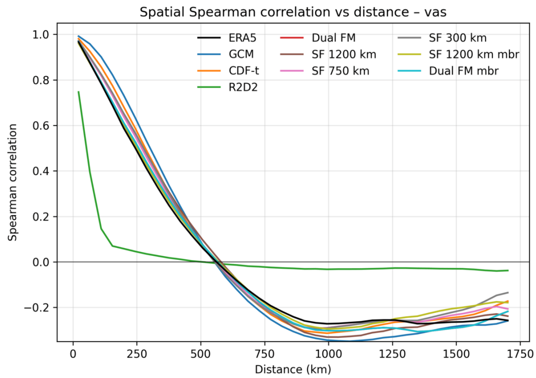
(a) Wind Speed



(b) Maximum wind speed

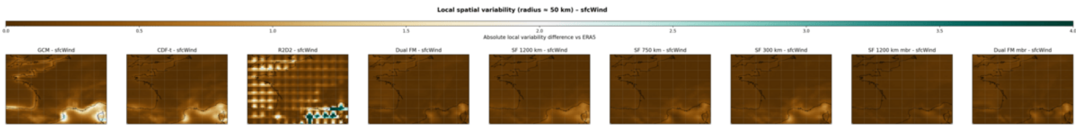


(c) Zonal wind

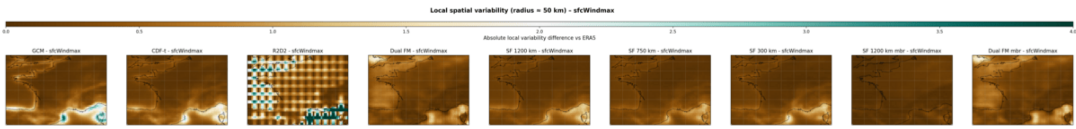


(d) Meridional wind

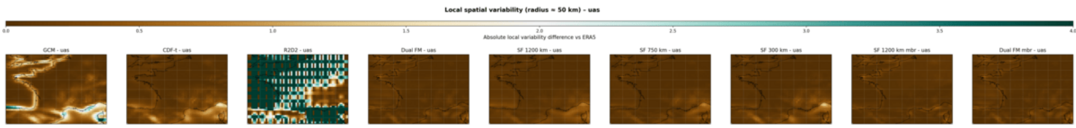
Figure 17. *Mean spatial Spearman correlation between pairs of locations as a function of their separation distance (distance bins in km) for each climate variable. Correlations are spatially averaged within each distance bin. On this plot, we can already see that the GCM grid points are much more correlated than those of ERA5. This is also the case for CDF-t, even though it is closer to the correlations of ERA5. While R2D2 is not too bad at reproducing wind speed correlations, it fails completely on u and v. Dual FM tends to decorrelate a little too strongly, and always ends up below the ERA5 curve. As for SerpentFlow, the curves are just above it. The lower the cutoff frequency, the closer we get to the ERA5 curve. The 1200 km single-member version fits the ERA5 curve almost perfectly for all variables.*



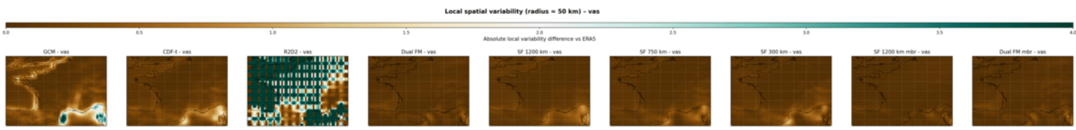
(a) Wind Speed



(b) Maximum wind speed



(c) Zonal wind



(d) Meridional wind

Figure 18. *Mean absolute difference in local spatial variability between each method and the ERA5 reference, computed over a neighborhood of approximately 50 km radius. Each panel corresponds to one method for a given climate variable. Values are averaged over all available times and grid points within the specified radius. The color scale indicates the magnitude of deviation from ERA5, with higher values representing stronger local differences. All maps show significantly higher biases on the sides and in mountainous areas (Alps and Pyrenees). However, the GCM shows much greater deviations over Corsica and the Mediterranean coast. For maximum wind speed, Dual FM significantly underperforms on the English coast and around Corsica. The single-member version of SerpentFlow 1200 km has much darker maps than the other methods, demonstrating its superior performance.*

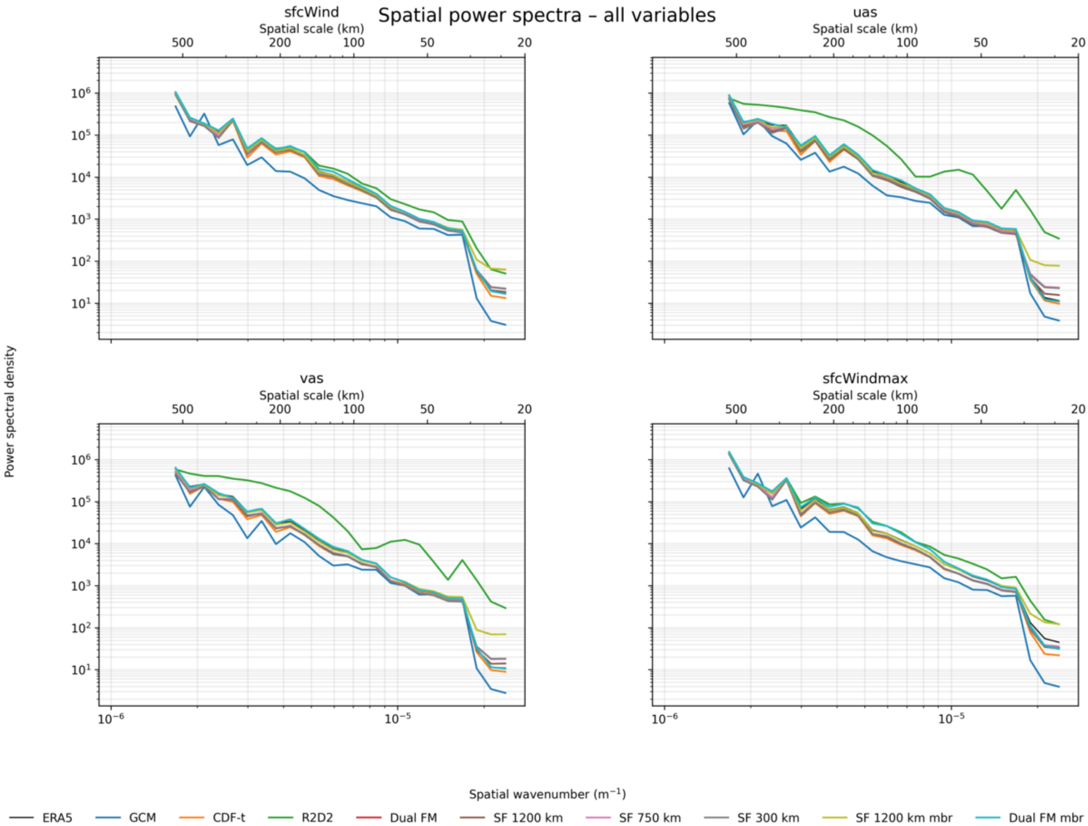


Figure 19. *Spatial power spectra for all variables. Each subplot shows the power spectral density as a function of spatial wavenumber for one variable. The bottom x-axis displays the wavenumber (m^{-1}), while the top x-axis indicates the corresponding spatial scale in kilometers. Spectra are computed over the full domain, and differences among methods reflect deviations from the reference across spatial scales. This figure shows that the GCM has much less energy at high frequencies (small-scale effects) compared to ERA5 and downscaling methods. All plots show a significant decrease in spectral energy for effects at less than 40 km. This decrease is much stronger for CDF-t and SerpentFlow cut at 300 and 750 km, particularly for sfcWindmax. Dual FM performs slightly better, but still falls short of SerpentFlow 1200 km.*

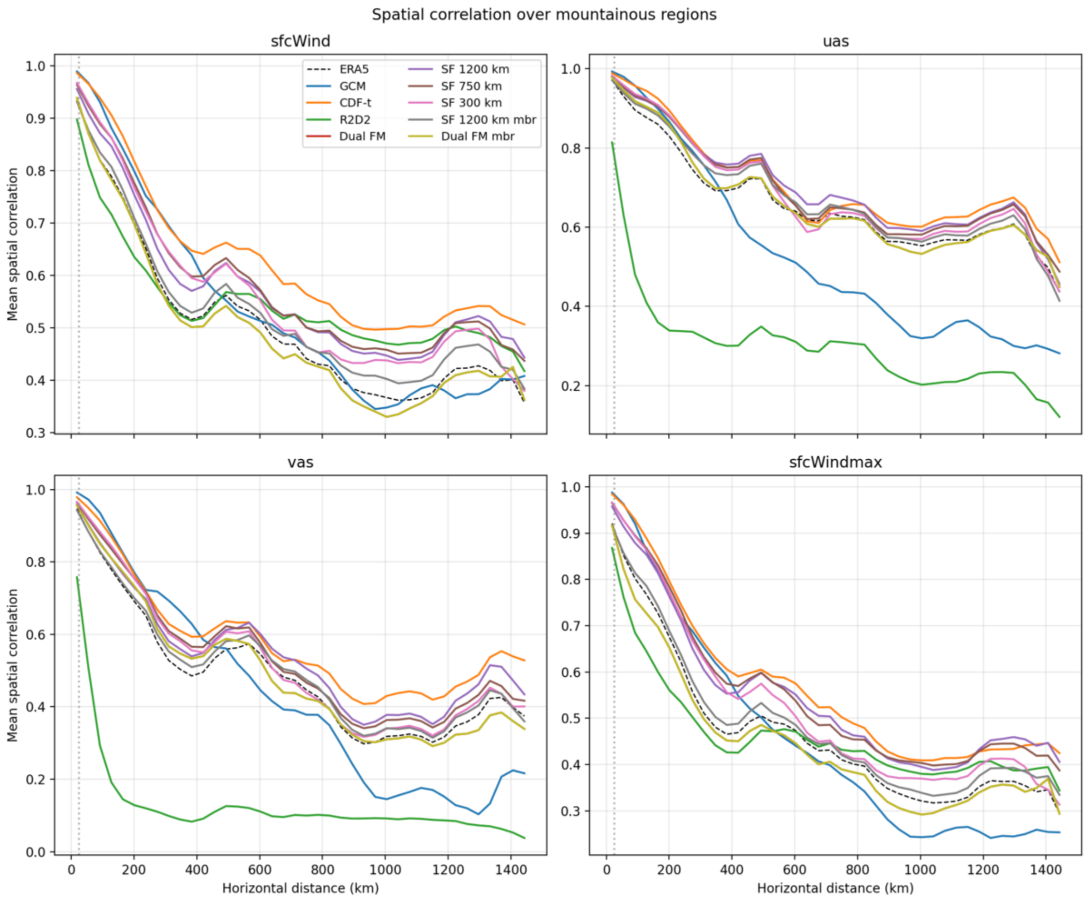


Figure 20. *Mean spatial correlations over mountainous regions (altitude > 800 m) as a function of horizontal distance. Each subplot corresponds to one variable. A vertical line at 25 km is added for reference. Correlations are computed only for grid points above 800 m to focus on high-elevation areas, highlighting method performance in orographic regions. The spatial over-correlations of the CDF-t and the GCM are much more visible in the mountains than in the plains (Figure 17). Here we can see that the lower the GCM frequencies are cut for SerpentFlow, the better we are at reconstructing the dynamics of ERA5. The SerpentFlow 1200 km average cuts out certain effects, making it slightly less accurate than the single-member version.*

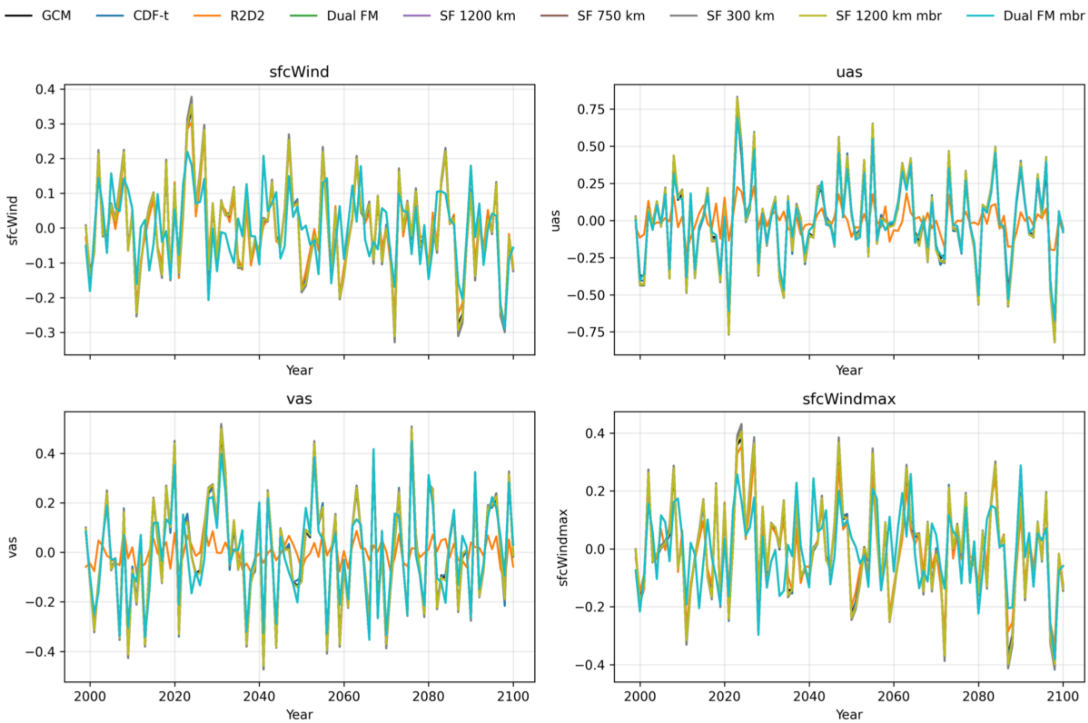


Figure 21. *Global annual anomalies for all variables. Each subplot shows one climate variable. With the exception of R2D2, which deviates for u and v, all methods follow the inter-annual variability of the GCM almost perfectly. As noted in (Keisler et al., 2026), Dual FM deviates from the GCM variations at certain peaks, predicting values that are too extreme.*

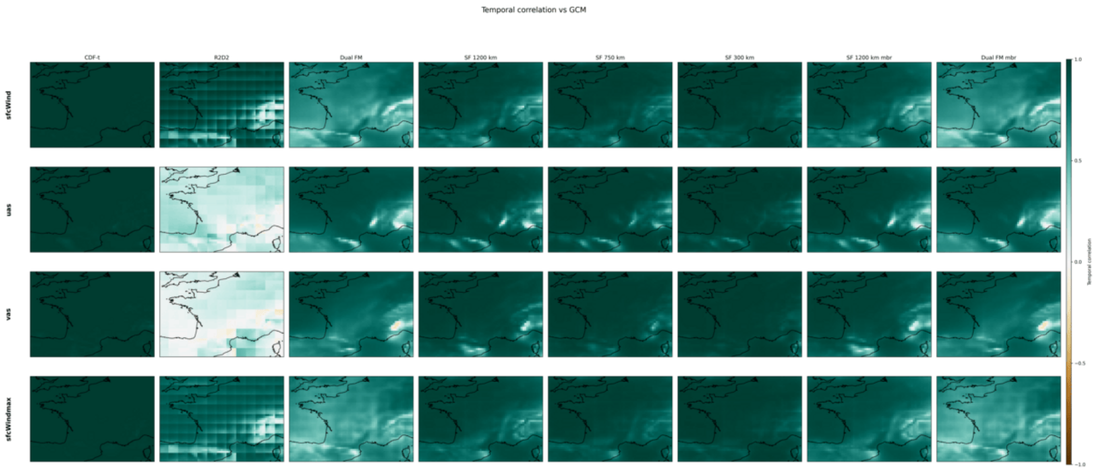
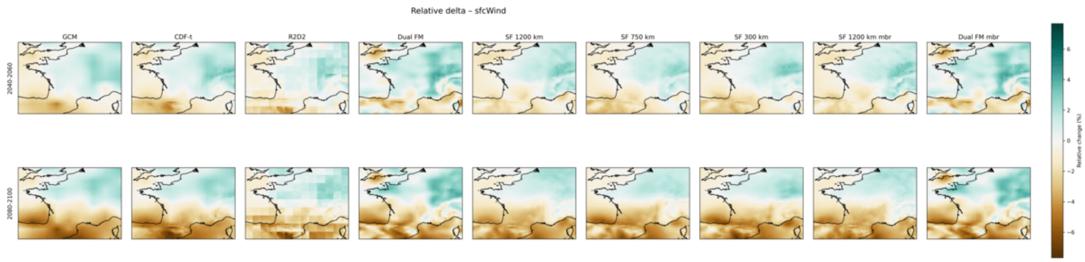
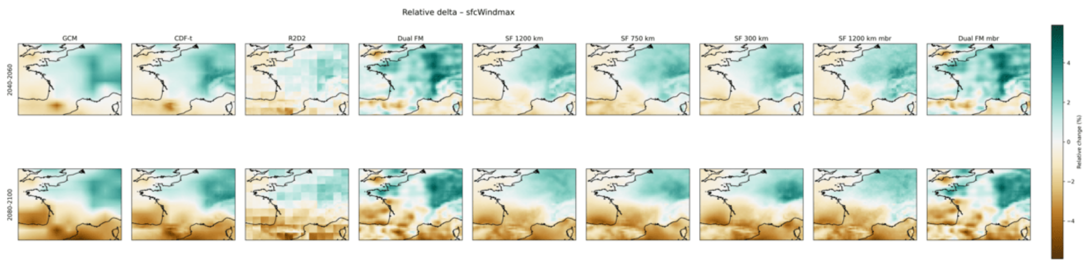


Figure 22. *Temporal correlation between each method and the reference GCM for all variables. Each row corresponds to one variable, and each column shows one method. The color scale indicates the correlation coefficient at each grid point, ranging from ranging from* -1 *(perfect anti-correlation) to* 1 *(perfect correlation). The CDF-t follows the temporal dynamics of the GCM grid point by grid point almost perfectly. Among the other methods, R2D2 falls completely behind for u and v, and Dual FM deviates more from the dynamics of the GCM than the various versions of SerpentFlow. The deviations are greater in mountainous areas (the Alps and Pyrenees) for deep learning methods. This is not surprising, as these are the regions where grid point values need to be modified the most to take into account local climatic effects and spatial correlations. It can be seen that the lower the SerpentFlow cut, the further it deviates from the GCM dynamics (and the closer it gets to the ERA5 dynamics).*



(a) Wind Speed



(b) Maximum wind speed

Figure 23. *Relative change (delta, in %) of each climate variable for multiple methods and future periods. Each row corresponds to a future period, and each column shows one method, including the reference GCM in the first column. Values represent the relative change with respect to the historical baseline, computed as % difference. Dual FM shows deltas that are quite significantly different from those of the GCM, with some increases in wind speed in Spain, whereas the GCM indicates a decrease. With regard to SerpentFlow, the lower the frequencies (e.g., for 1,200 km), the more local effects of increases or decreases emerge that are not present in the GCM. It is difficult to know whether these local effects are plausible or not, as they cannot, by design, be modeled by the GCM, which only represents effects on a larger scale.*

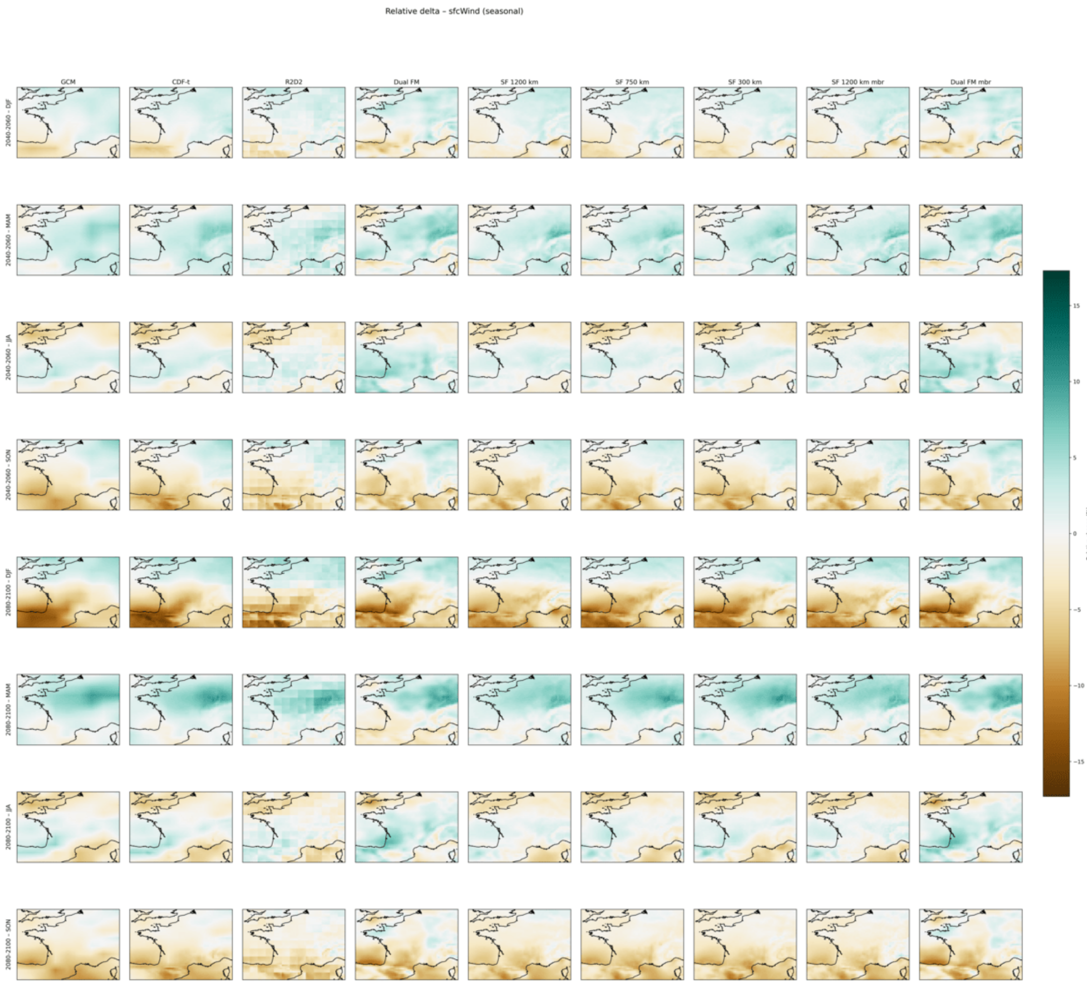


Figure 24. *Relative seasonal change (delta, in %) of mean wind speed for multiple methods and future periods. Each row corresponds to a future period and a season, and each column shows one method, including the reference GCM in the first column. Values represent the relative change with respect to the historical baseline, computed as % difference. Results are similar to the ones from Figure 23.*

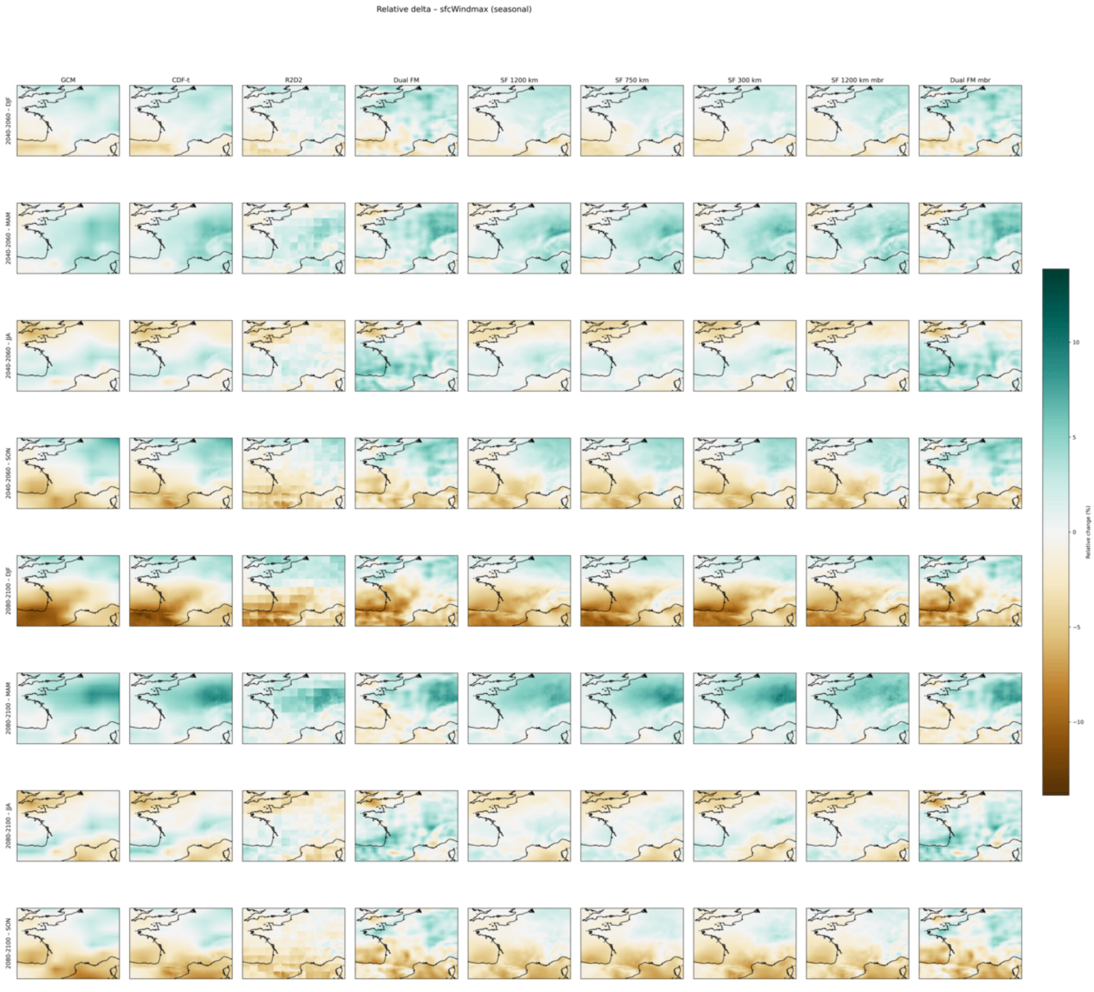
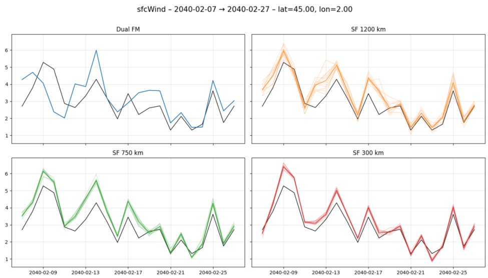
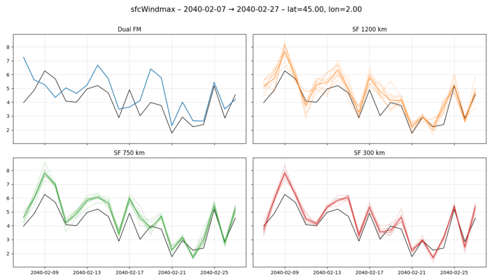


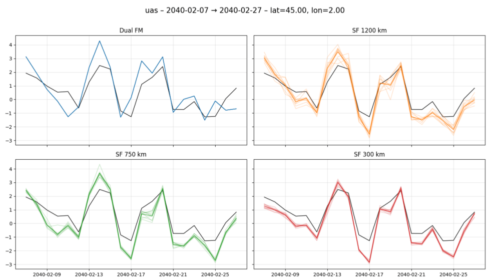
Figure 25. *Relative seasonal change (delta, in %) of maximal wind speed for multiple methods and future periods. Each row corresponds to a future period and a season, and each column shows one method, including the reference GCM in the first column. Values represent the relative change with respect to the historical baseline, computed as % difference. Results are similar to the ones from Figure 23.*



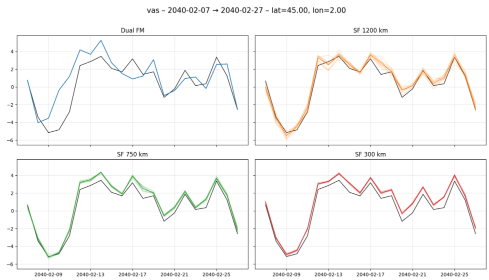
(a) Wind Speed



(b) Maximum wind speed



(c) Zonal wind



(d) Meridional wind

Figure 26. *Time series of each climate variable at a specific location (lat=45, lon=2) over a 20-day window starting 2040-02-07. Each panel shows one method: the colored line is the ensemble mean, faint colored lines show individual ensemble members, and the black line represents the GCM reference. For Dual FM, all ensemble members are identical, resulting in a single trajectory that does not closely follow the GCM dynamics; this demonstrates that, despite its generative design, the method is effectively deterministic and does not capture uncertainty. In contrast, SerpentFlow's generative framework allows for variability: a smaller frequency cutoff (1200 km) produces a wider spread that captures more local variability but may deviate further from the GCM, whereas larger cutoffs (300 km) constrain the model more, leading to a tighter spread and closer agreement with the reference. From a climate perspective, this reflects the trade-off between representing local variability and reproducing the large-scale climate signal: a wider spread better represents the range of possible values at each point while maintaining the observed spatial and temporal structure, whereas a tighter spread prioritizes fidelity to the reference climatology..*

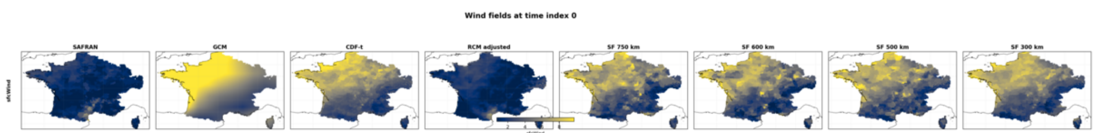


Figure 27. *Wind maps for the first time step, for each method. The reanalysis values are only available for mainland France and Corsica. The reanalysis and downscaling methods (with the exception of the RCM) present small spatial patches, due to the design of the reanalysis itself. The data between SAFRAN and el GCM are not aligned in time, which explains the large difference between the two. The RCM represents a climate that is still different from the GCM, which explains the differences in the maps. We can see that the SerpentFlow (SF) versions are all slightly different. The lower the cutoff (300 km or 500 km), the closer we get to the CDF-t.*

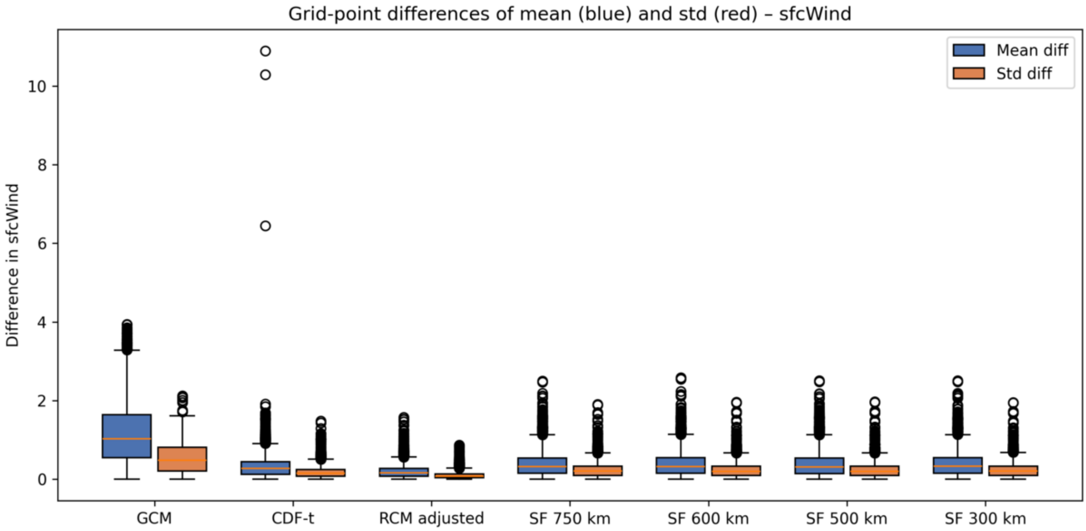


Figure 28. *Distribution of mean and standard deviations differences w.r.t SAFRAN per grid point. Compared to the bias reduction between ACCESS and ERA5 (see Figure 9), this reduction is less pronounced, even for the CDF-t, which also shows a few rare large deviations from the mean. The performance of the different versions of SerpentFlow is fairly similar.*

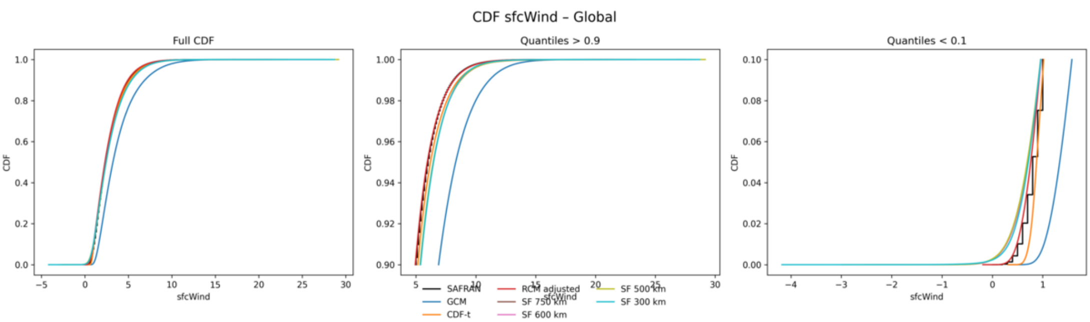


Figure 29. *Cumulative Distribution Functions (CDFs) for each method over the validation period. For each sublot a zoom is done on the extreme quantiles ($q \leq 0.1$ and $q \geq 0.9$). SAFRAN's CDF is discontinuous in places due to the rounding used for its values, which may also explain the discrepancies seen in Figure 28. Apart from the RCM, which is very close to the reanalysis, the other methods tend to overestimate wind speed (still much less than the GCM). SF 300 km produces some negative values.*

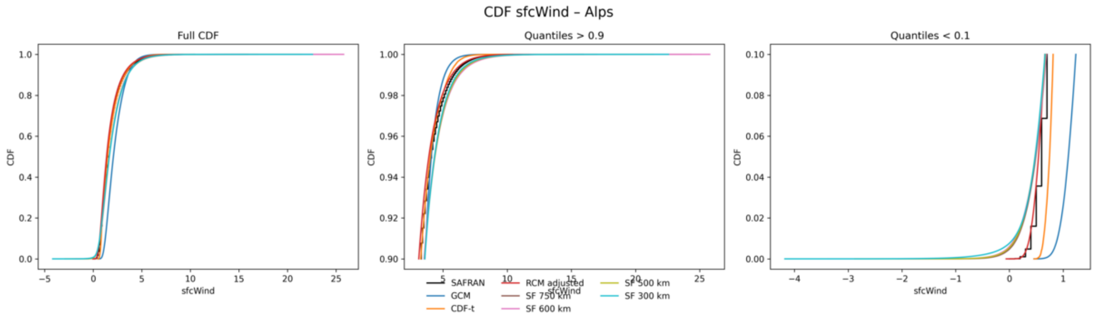


Figure 30. *Cumulative distribution functions (CDFs) for each method over a small region in the Alps (*latitude $\in [44, 47]$, longitude $\in [5, 8]$*). For each subplot, a zoom is applied to the extreme quantiles (*$q \leq 0.1$ *and* $q \geq 0.9$*). The distribution correction is also correct on a small regional scale. In the high quantiles, the t-CDF tends to follow the GCM and underestimate the highest values. .*

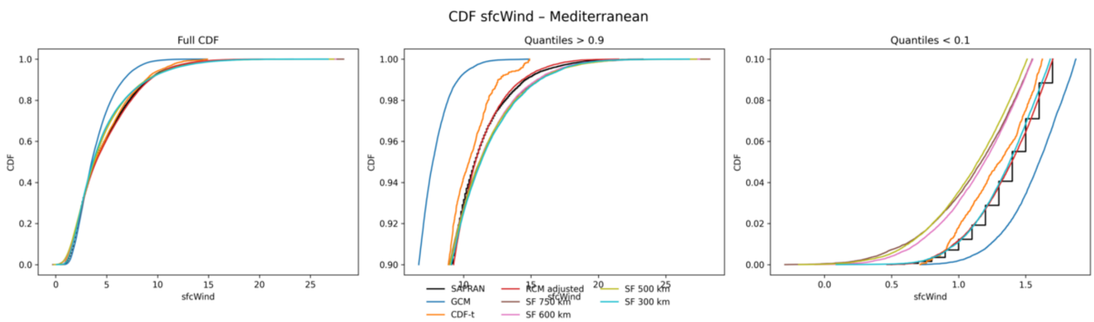


Figure 31. *Cumulative distribution functions (CDFs) for each method over a small region in the Mediterranean (*latitude $\in [41, 43]$, longitude $\in [3, 6]$*). For each subplot, a zoom is applied to the extreme quantiles (*$q \leq 0.1$ *and* $q \geq 0.9$*). In this region, bias correction is slightly less effective for all methods except RCM. CDF-t is unable to reproduce the correct values at all for the high quantiles.*

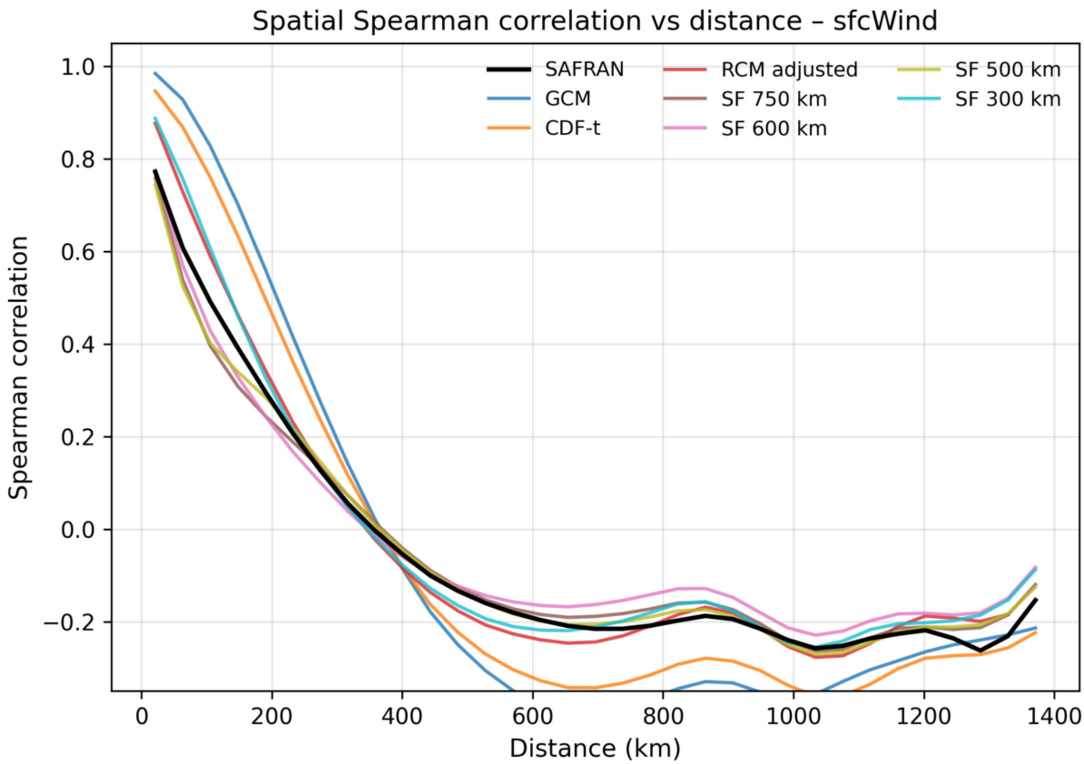


Figure 32. *Mean spatial Spearman correlation between pairs of locations as a function of their separation distance (distance bins in km) for each climate variable.Correlations are spatially averaged within each distance bin. It is interesting to note here that SerpentFlow (and in particular the 500 km version) is better than RCM at modeling the spatial dynamics of SAFRAN.*

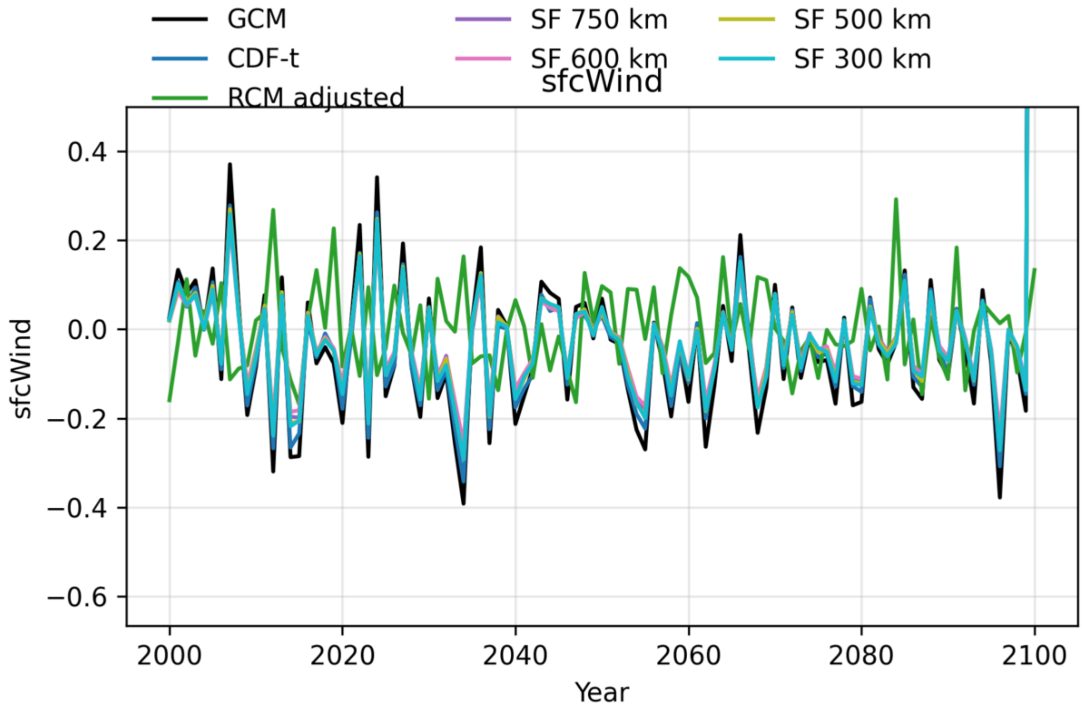


Figure 33. Global annual anomalies. Since the RCM models a different climate than the GCM, it is quite normal for the interannual dynamics to be different. As for the other methods, while they generally follow the GCM signal very closely, they tend not to reach the most extreme values.

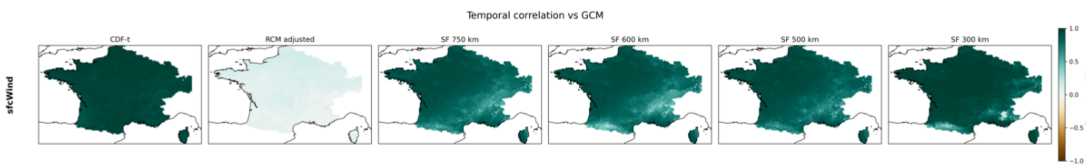


Figure 34. *Temporal correlation between each method and the reference GCM. Each column shows one method. The color scale indicates the correlation coefficient at each grid point, ranging from ranging from -1 (perfect anti-correlation) to 1 (perfect correlation). As shown in Figure 33, the RCM does not follow the dynamics of the GCM. For SerpentFlow, as seen in the ERA5 example (see Figure 22), the more we cut at low frequencies, the further we move away from the dynamics of the GCM, particularly over mountainous terrain (the Alps and Pyrenees).*

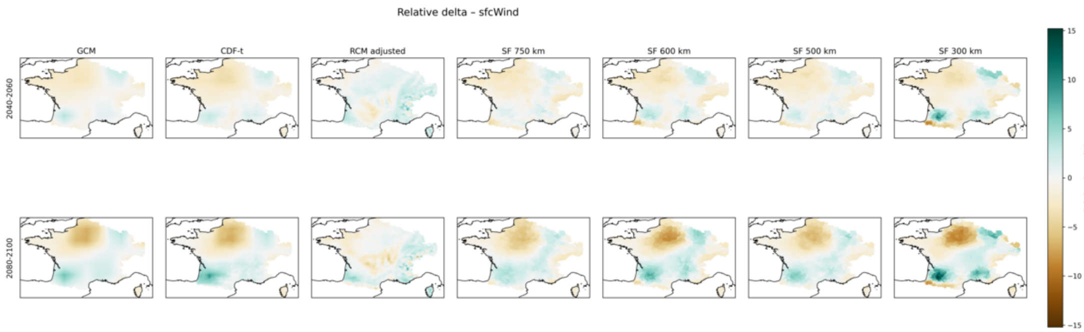


Figure 35. *Relative change (delta, in %) for multiple methods and future periods. Each row corresponds to a future period, and each column shows one method, including the reference GCM in the first column. Values represent the relative change with respect to the historical baseline, computed as % difference. It is interesting to note that even for these deltas, the signal proposed by the RCM is quite different from that of the GCM. While we would expect the signal to be quite different from year to year, we might have expected a similar overall trend, as both models simulate the same scenario. However, we can see that the decrease in northwestern France, for example, is much less pronounced for the RCM than for the GCM. Otherwise, the dynamics of the SF versions are fairly similar to those of the GCM, with more pronounced variations for the SF 300 km version..*

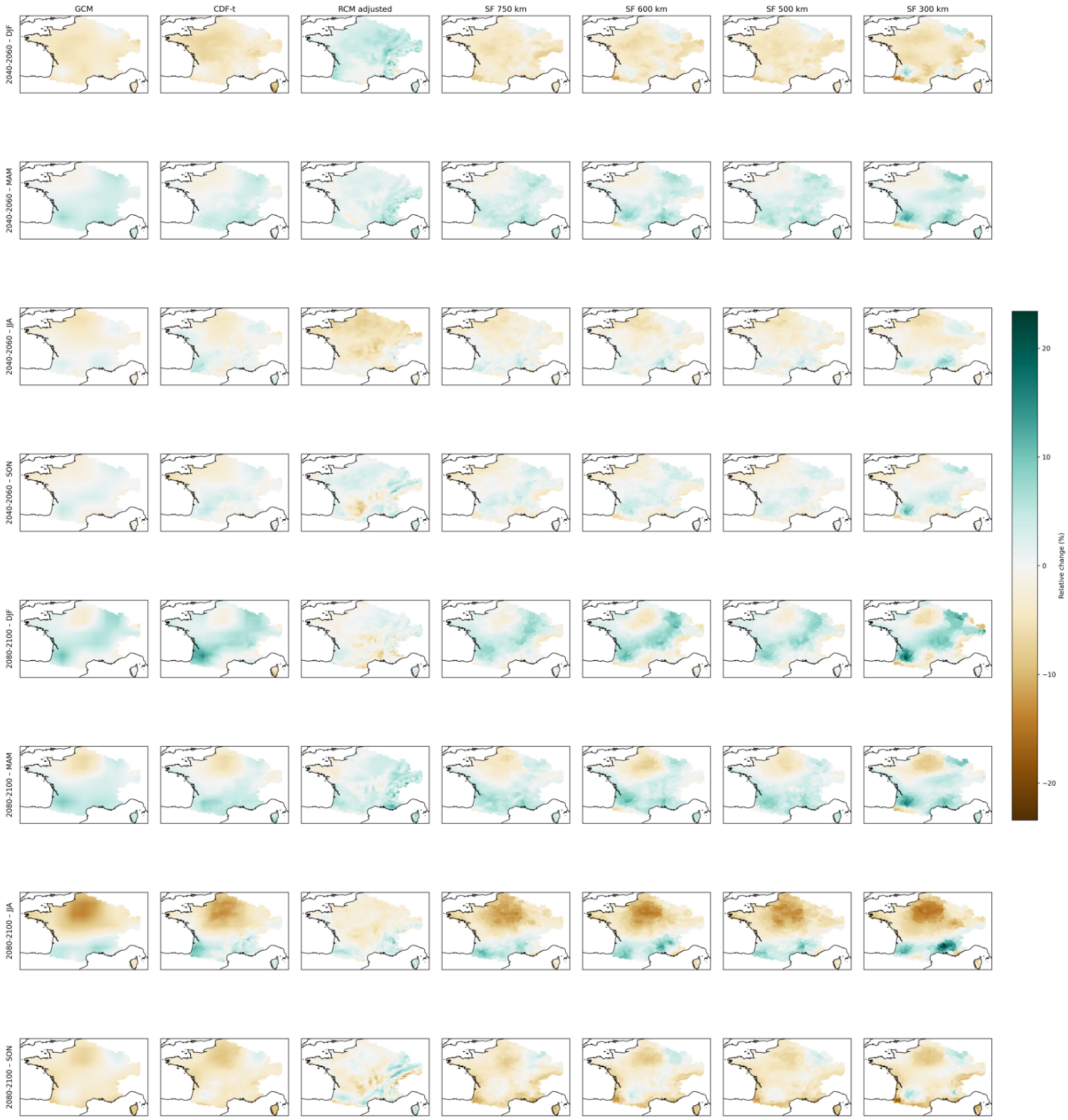


Figure 36. *Relative seasonal change (delta, in %) for multiple methods and future periods. Each row corresponds to a future period and a season, and each column shows one method, including the reference GCM in the first column. Values represent the relative change with respect to the historical baseline, computed as % difference..*