Supplement to

Evaluation of nine gridded daily weather reconstructions for the European heatwave summer of 1807

Peter Stucki[1,2], Stefan Brönnimann[1,2], Noemi Imfeld[1,2], Lucas Pfister[1,2], Conall E. Ruth[1,2], Yannis Schmutz[1,2,3], Yuri Brugnara[1,2], Martin Wegmann[1,2], Rajmund Przybylak[4,5], Janusz Filipiak[6]

[1] Oeschger Centre for Climate Change Research, University of Bern, Bern, 3012, Switzerland
[2] Institute of Geography, University of Bern, Bern, 3012, Switzerland
[3] Berner Fachhochschule Technik und Informatik, Bern, 3012, Switzerland
[4] Faculty of Earth Sciences and Spatial Management, Nicolaus Copernicus University, Toruń, Poland
[5] Centre for Climate Change Research, Nicolaus Copernicus University, Toruń, Poland
[6] Department of Physical Oceanography and Climate Research, Faculty of Oceanography and Geography, University of Gdańsk, Gdańsk, Poland

Correspondence to: Peter Stucki (peter.stucki@unibe.ch)

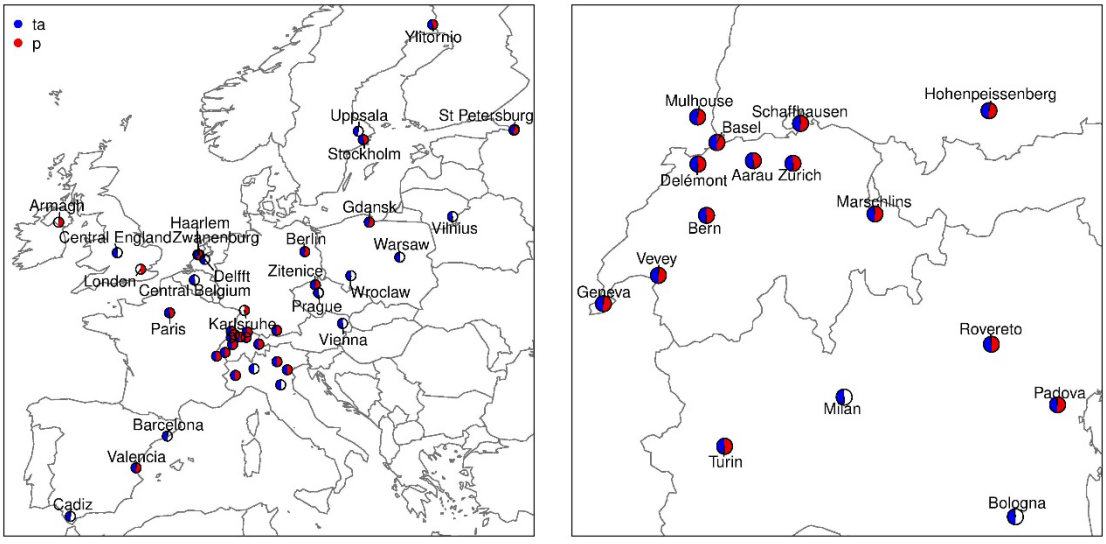

Figure S1: Set of early instrumental measurements considered for this study. Half circles mark stations with measurements of 2-meter air temperature (ta; blue) and sea level pressure (p; red). Note that not all stations have observations for each analysis in the main text.

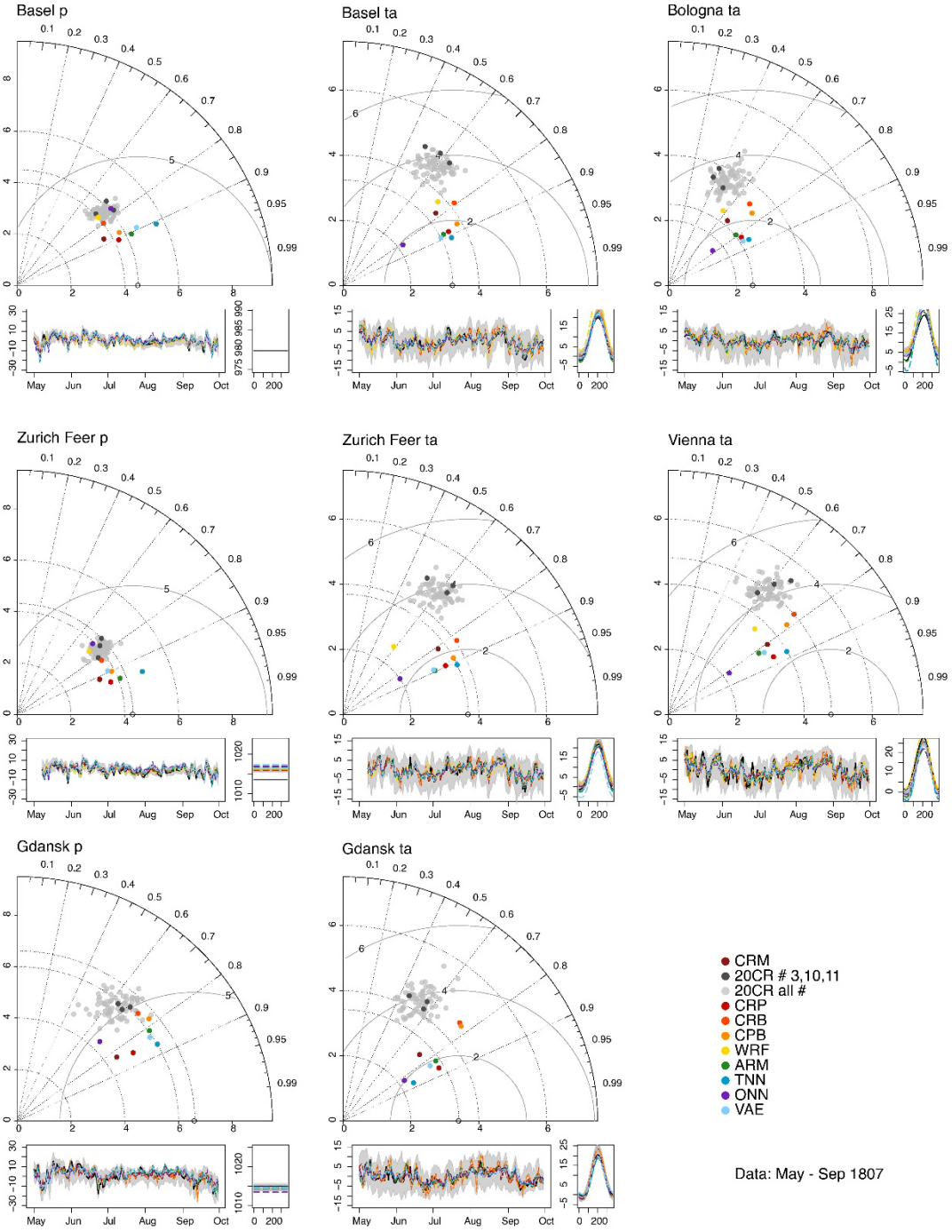

Figure S2: Taylor diagrams (top area in each panel), time series (bottom left area in each panel), and climatology baseline (bottom right area of each panel; fit from the combined first two harmonic waves for temperature, annual mean for pressure) of observed

(black circle or black bold line) versus predicted (colored dots or dashed lines; see image legend for abbreviations) anomalies of 2-meter air temperature (ta) and pressure (p) for the locations of Basel, Bologna, Zurich, Vienna, and Gdansk (from top left to bottom right) in the period May to September 1807.

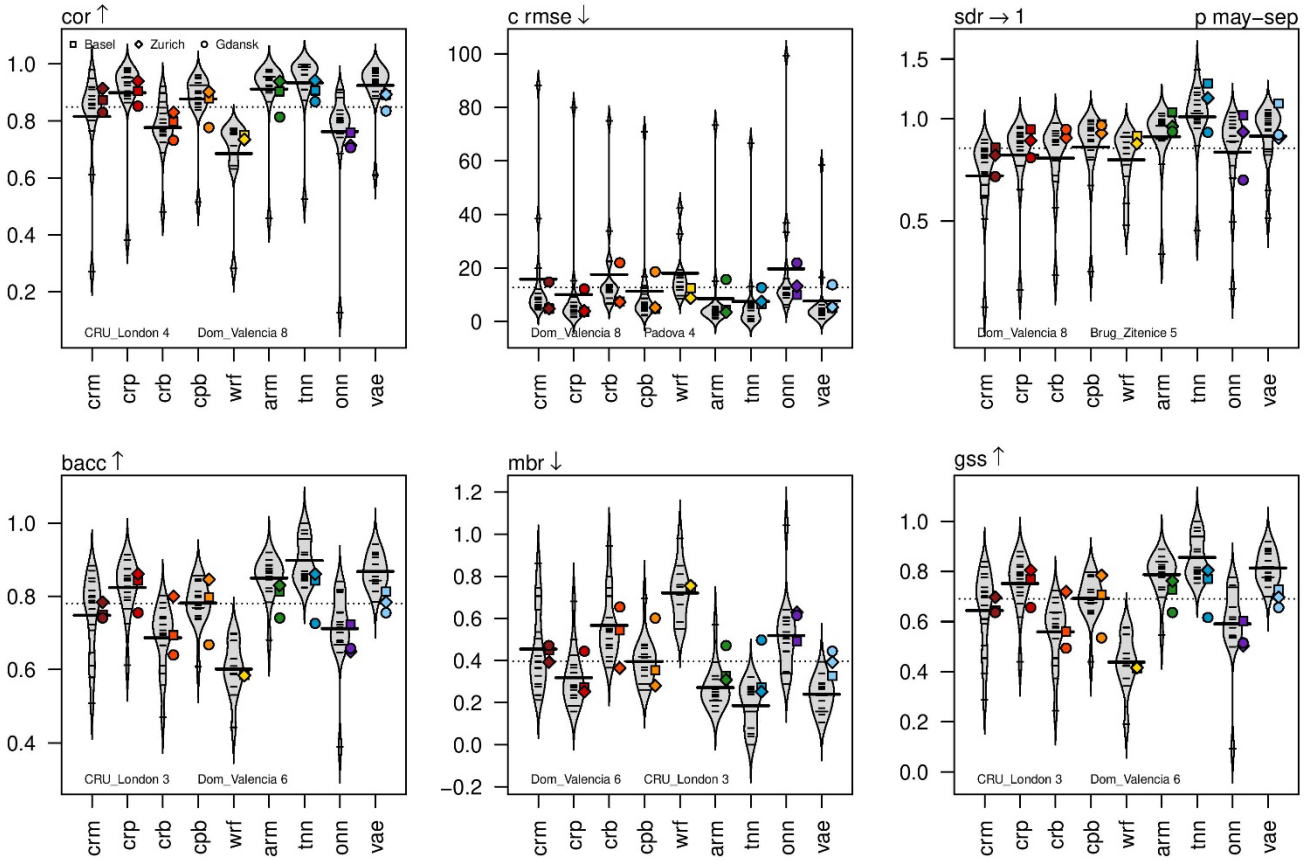

Figure S3: As in Figure 5 in the main text, but for pressure.

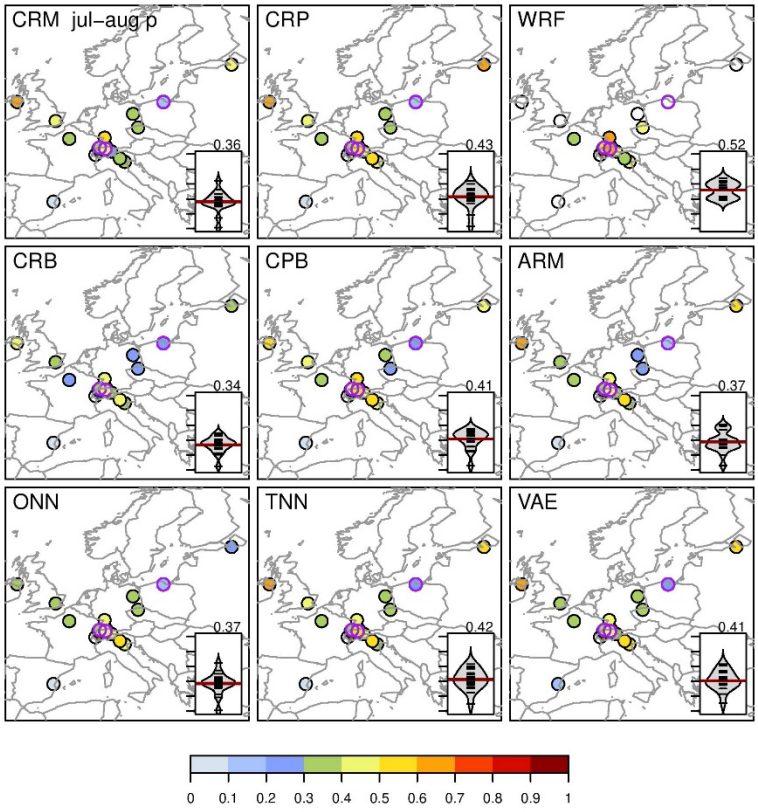

Figure S4: As in Figure 6, but for pressure anomalies.

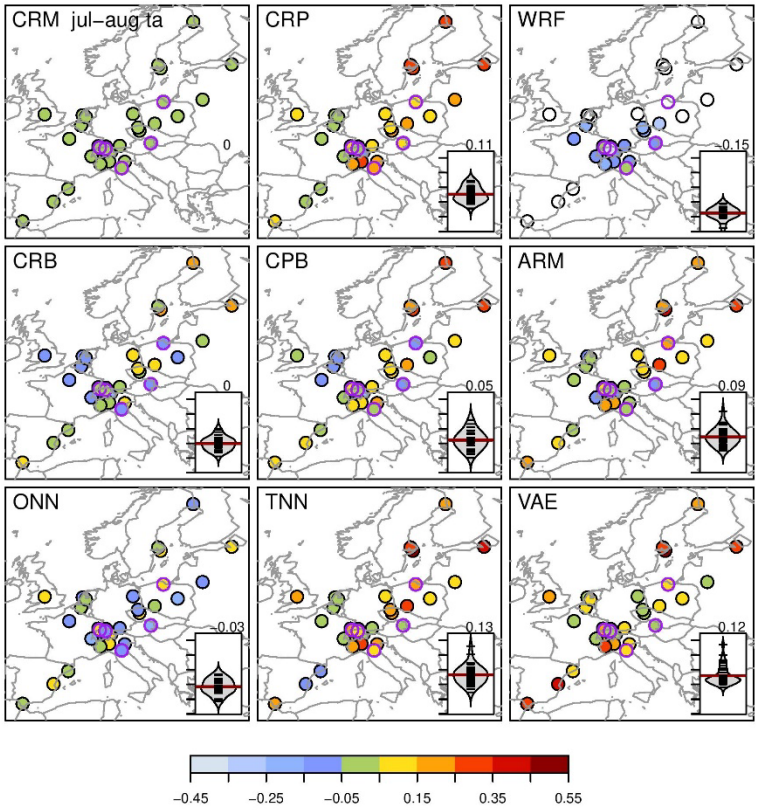

Figure S5: As in Figure 6, but showing the difference in *TSS* (temperature anomalies) with regards to CRM.

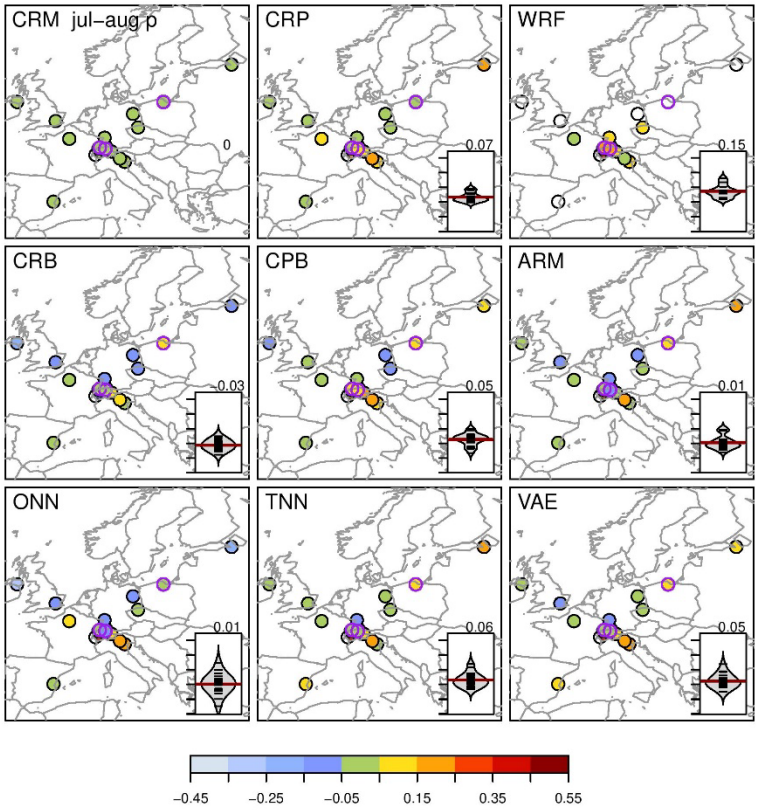

Figure S6: As in Figure 6, but showing the difference in *TSS* (pressure anomalies) with regards to CRM.

40

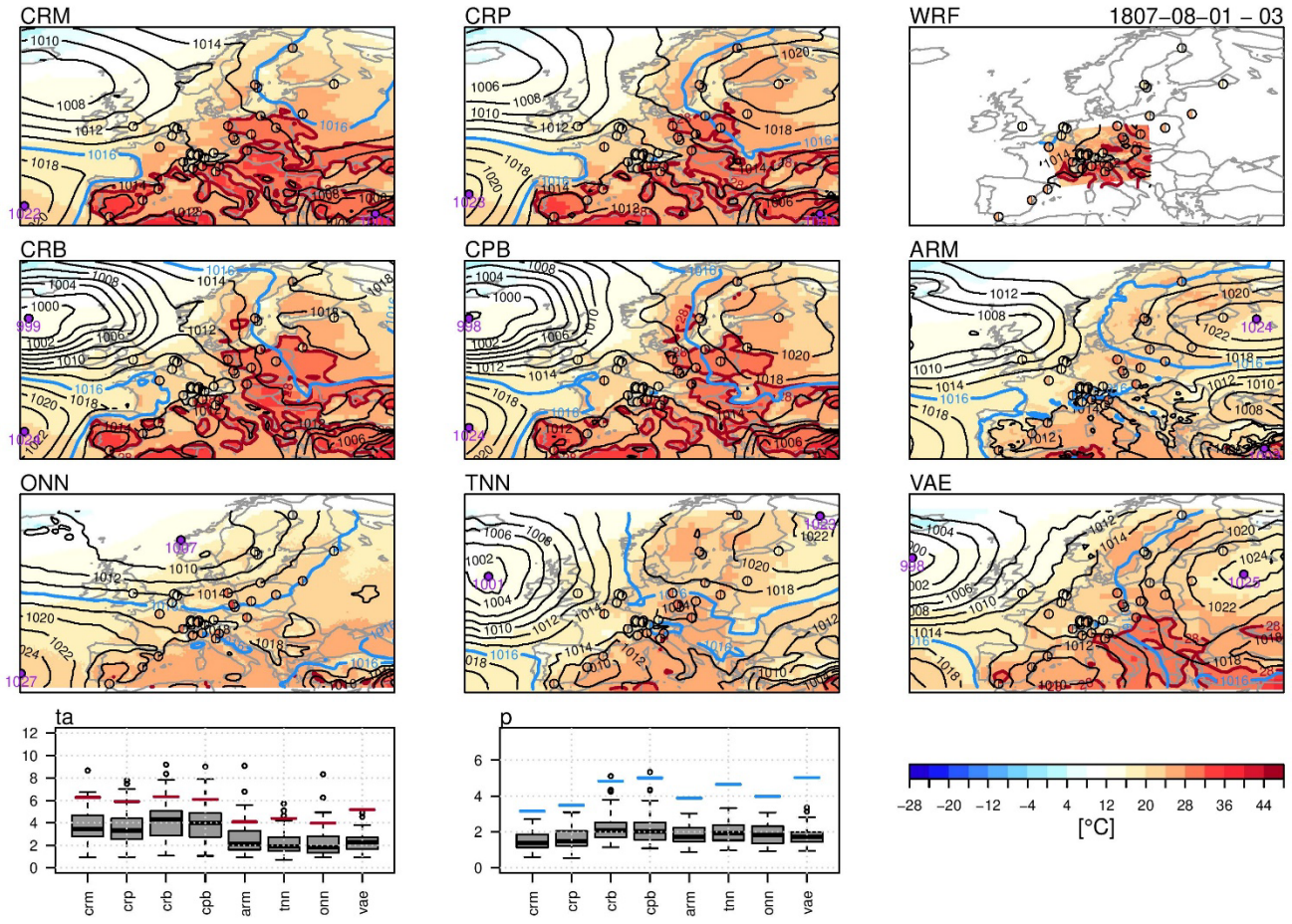

Figure S7: As in Figure 7, but for the mean state in the period 1 to 3 August 1807.

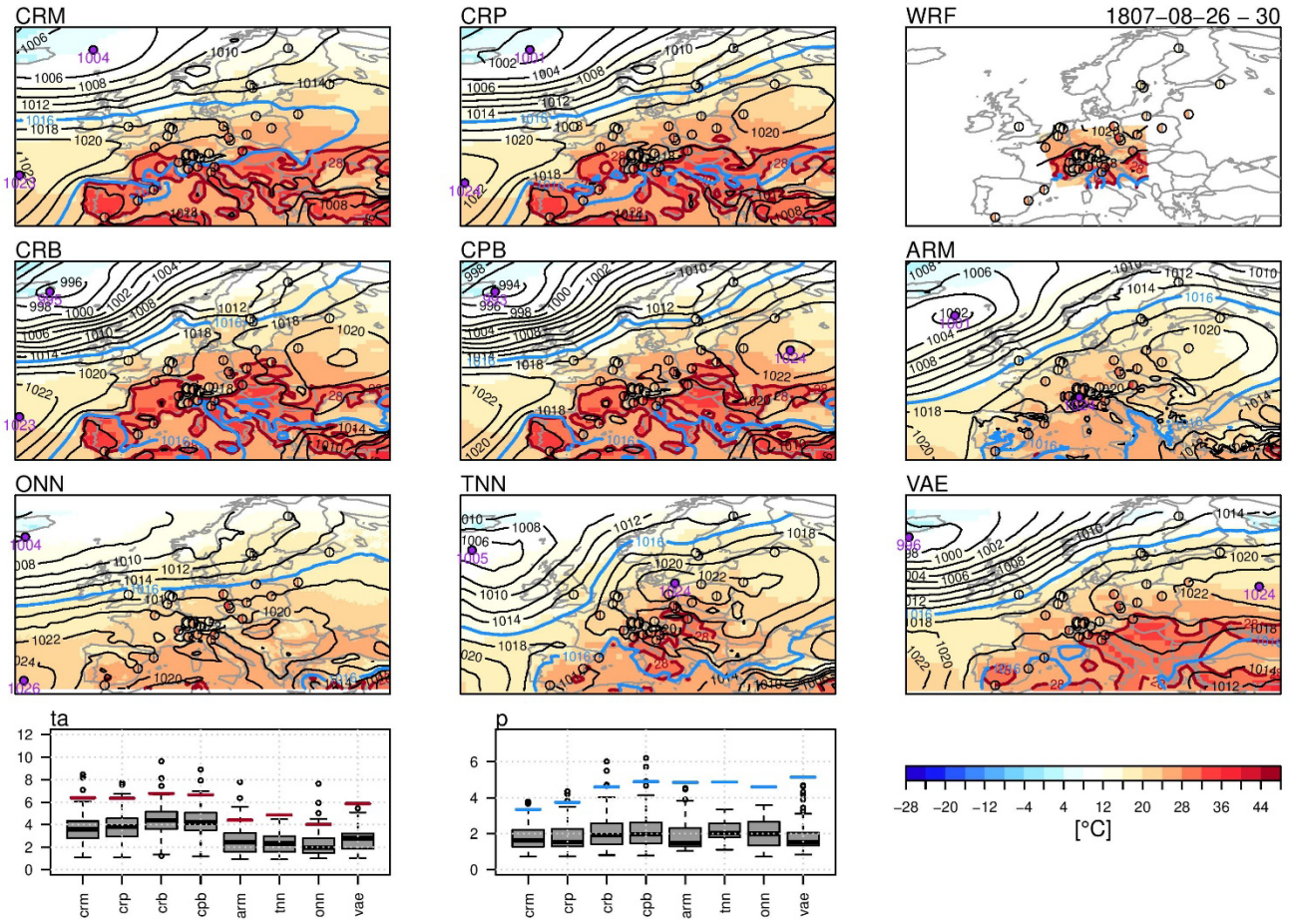

Figure S8: As in Figure 7, but for the mean state in the period 26 to 30 August 1807

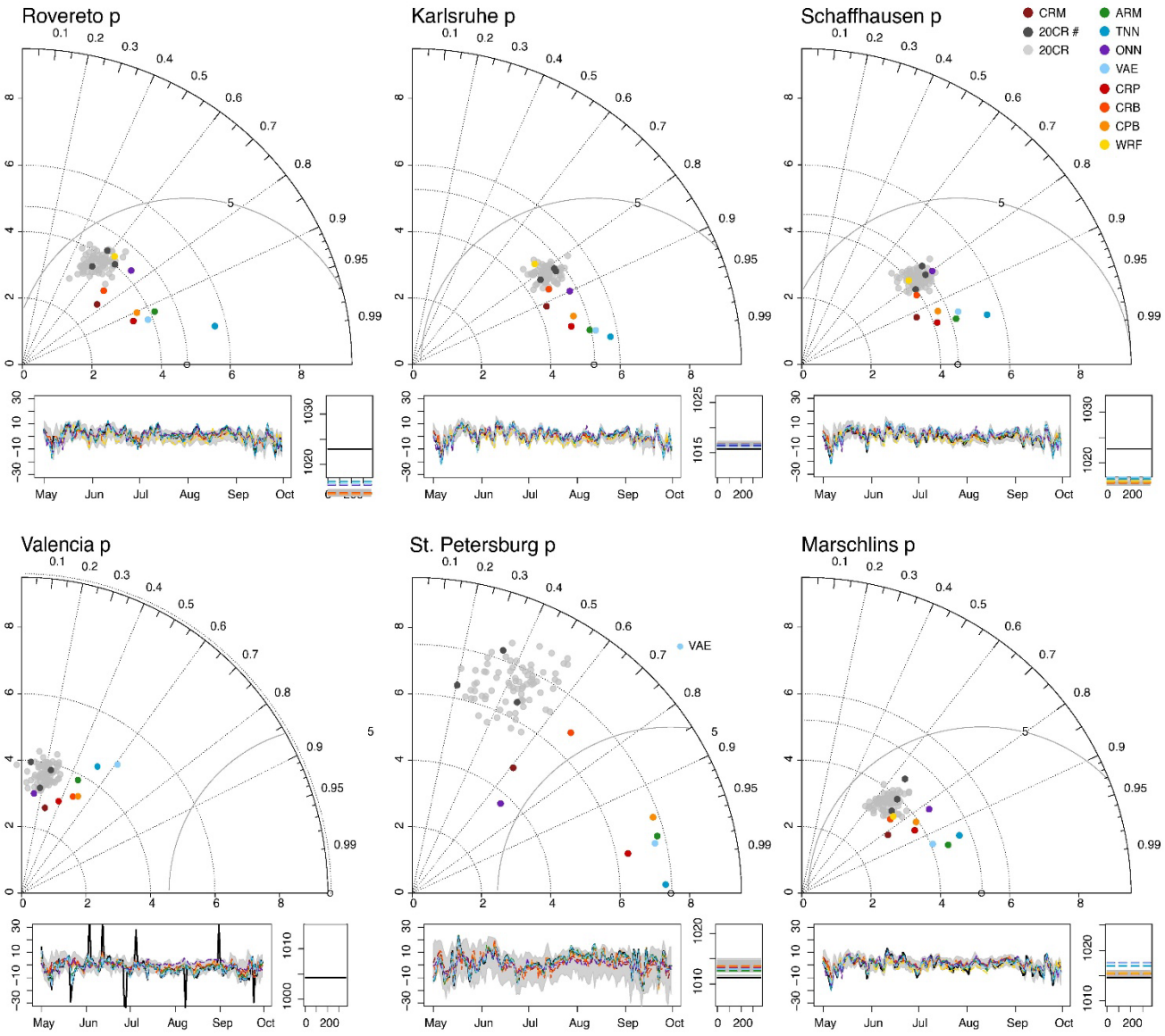

Figure S9: As in Figure S2 and Figure 9, but for different stations and pressure (anomalies wrt annual mean).

S.1 Variational auto-encoder: Supplemental details

S.1.1 Model architecture

A detailed description of the VAE architecture is given in **Table S1**. The model takes an input of size 32 * 64 * 2 (corresponding to: latitude * longitude * temperature/pressure) and compresses it into a 256-dimensional latent space. Convolutional layers in the encoder extract spatial features while progressively downscaling the input. Near the end of the encoder, a fully connected layer transforms the downscaled feature maps into two 256-dimensional vectors, representing the mean and log-variance of the latent distribution. To enable stochastic sampling while maintaining differentiability, the 'sampling layer' applies the "reparameterisation trick" of Kingma & Welling (2014), ensuring that the model remains trainable using gradient-based optimisation. The decoder then reconstructs the input from the sampled latent vector, first mapping it to a 4 * 8 * 160 feature projection in the 'project and reshape layer'. Transposed convolutional layers progressively upscale the spatial dimensions, generating coarse reconstructions of the original input, while additional convolutional layers are included to refine spatial details, smooth feature maps, and ensure a more accurate reconstruction. The final layer of the decoder adjusts the output to a size of 32 * 64 * 2, matching the original input.

Table S1: Architecture of the variational auto-encoder. Down and up arrows after the layer type indicate a halving and doubling of the spatial size, respectively.

	Layer type	Kernel	Filters	Padding/ cropping	Stride	Additional info
Encoder	Image input	-	-	-	-	Input size: 32*64*2
	2D convolution	3	20	same	1	Batch normalisation; ReLU activation; Dropout rate: 0.05
	2D convolution (↓)	5	40	same	2	Batch normalisation; ReLU activation; Dropout rate: 0.05
	2D convolution (↓)	5	80	same	2	Batch normalisation; ReLU activation; Dropout rate: 0.1
	2D convolution (↓)	7	160	same	2	Batch normalisation; ReLU activation; Dropout rate: 0.15
	2D convolution	7	160	same	1	Batch normalisation; ReLU activation; Dropout rate: 0.25
	Fully connected	-	-	-	-	Output size: 512 (mean and log variance of each latent variable)
	Sampling	-	-	-	-	Uses "reparameterisation trick"
Decoder	Feature input	-	-	-	-	Input size: 256
	Project and reshape	-	-	-	-	Projection size: 4*8*160
	2D convolution	7	160	same	1	ReLU activation
	2D transposed convolution (↑)	7	80	same	2	ReLU activation
	2D transposed convolution (↑)	5	40	same	2	ReLU activation
	2D transposed convolution (↑)	5	20	same	2	ReLU activation
	2D convolution	3	10	same	1	tanh activation
	2D transposed convolution	3	2	same	1	tanh activation
	Linear	-	-	-	-	Output size: 32*64*2

S.1.2 Data preparation

To maximise the comparability with TNN, the VAE is based on ERA5 fields of 2-meter air temperature and MSLP for the period 1950-2020, aggregated to mean daily values at a 1° resolution for the domain 36N-67N, 22W-41E. Here, the data are divided into a training set (1969-2020), primary validation set (1962-1968), secondary validation set (1955-1961) and test set (1950-1954). The training and primary validation sets relate to the set-up of the entire auto-encoder, whereas the secondary validation and test sets relate to the application of the trained decoder for reconstructing partially observed weather fields. To prepare the data, each grid cell is treated individually. For temperature, the observed mean value of each calendar month is subtracted such that trends and inter-monthly seasonality are removed while intra-monthly seasonality is retained. For pressure, any trends or seasonality are assumed to be negligible relative to the day-to-day variability, therefore the overall mean of the training period is subtracted instead. The values of each variable are then divided by the (resulting) standard deviations of the training period. The 1807 series are prepared similarly using the observed monthly mean temperature, and the mean pressure and standard deviations of the corresponding aggregated grid cell. If necessary for estimating the monthly means, a provisional gap filling of temperature is performed by fitting the first two harmonics within a 5-year moving window; any infilled values are subsequently discarded. To facilitate the model evaluation, the secondary validation and test sets are each prepared assuming full availability and assuming the availability of 1807.

S.1.3 Training procedure

The VAE's overall training loss function, L, is given by:

$$L = \alpha * L_{rec} + \beta * L_{KL} ,$$

where L_{rec} is the mean squared difference between the model input and output, L_{KL} is the Kullback-Leibler divergence, which regularises the distribution of each latent variable to align with a standard normal distribution (Kingma & Welling, 2014), and α and β are constants tuned to 1.5 and 0.5, respectively. Including L_{KL} is crucial to ensure a well-structured latent space allowing the generation of new samples. The batch size is tuned to 64, and the learning rate is tuned to 5e-4, decaying by 7.5% after each epoch. The model is trained with a patience of 10 epochs (with respect to the primary validation set), resulting here in convergence after 42 epochs (52 epochs completed in total). The training procedure took approximately 3 hours on a standard (CPU) laptop.

S.1.4 Application procedure

To apply the trained decoder to reconstruct the partially observed weather fields for a given day, the latent space is randomly initialised and then iteratively adjusted up to 500 times such that the mean squared error between the output and any observed grid cells is minimised. The adjustment is based on gradient descent with an associated learning rate tuned to 0.1. On a standard

(CPU) laptop, reconstructing one year of daily fields in this manner took approximately 1 hour 40 minutes. After reconstructing the entire period of interest, a simple post-processing adjustment is applied to the temperature values to remove any net anomaly within each calendar month. For both temperature and pressure, the values are converted from normalised to absolute anomalies by multiplying by the standard deviation of the given grid cell within the training set. The reconstructions are then converted to fully absolute terms by adding the training-set means of pressure and the monthly re-analysis temperature values of ModE-RA (Valler et al., 2024).

S.1.5 Tuning and evaluation

The VAE's exact architecture and hyper-parameters have been tuned to minimise the RMSE between the secondary validation set and the corresponding complete reconstructed fields (assuming the observation availability of 1807). The performance is then re-evaluated on the test set, yielding overall RMSEs of 2.52 K and 5.18 hPa. We find that the model performs relatively well over central Europe, although it exhibits more pronounced difficulties reconstructing sparsely observed regions.

References

Brohan, P.: Data Assimilation with a deep convolutional variational autoencoder, ECMWF–ESA Workshop on Machine Learning for Earth Observation and Prediction, Reading 14–17 November 2022.

Kingma, D. P., and Welling, M.: Auto-encoding variational Bayes, In Conference proceedings: Papers accepted to the International Conference on Learning Representations (ICLR), 2014.

Valler, V., et al.: ModE-RA: A global monthly paleo-reanalysis of the modern era 1421 to 2008, Nature: Scientific Data, 11:36, https://doi.org/10.1038/s41597-023-02733-8, 2024.