

Physically-Informed Video Inpainting: A Deep Learning Approach for Historical Weather Reconstruction

Yannis Schmutz¹, Noemi Imfeld^{2,3}, Stefan Brönnimann^{2,3}, Erik Graf¹

¹Applied Machine Intelligence, Bern University of Applied Sciences, Bern, Switzerland

²Oeschger Center for Climate Change Research, University of Bern, Bern, Switzerland

³Institute of Geography, University of Bern, Bern, Switzerland

Key Points:

- This paper proposes a novel deep-learning based video-inpainting method for the reconstruction of daily historical weather fields.
- The implementation of domain-specific modeling improvement techniques halves the validation error.
- Our model reconstructs the heat wave of 1807 in Europe with a high degree of accuracy despite 99% missing cells.

Corresponding author: Yannis Schmutz, yannis.schmutz@bfh.ch

Abstract

We investigate the applicability of deep learning methods for reconstructing daily weather data. Inspired by video inpainting, we propose a novel method, WeRec3D, which utilizes a three-dimensional convolutional neural network. Our approach was developed iteratively by evaluating six modeling improvement techniques. The resulting method reduces the validation error to 48% compared to the baseline. Additionally, we demonstrate the impact of the spatial distribution of observations on reconstruction accuracy and propose a potential integration with the analogue resampling method. WeRec3D is trained and validated in a self-supervised manner using ERA5’s surface temperature and pressure data over Europe. On a hold-out set from 1950 to 1954, the validation results in an MAE of 1.11 °C and 199 Pa. As a case study, we reconstruct the 1807 heat wave and validate it using a leave-one-out method in space. Compared to the original data, the reconstructed time series exhibit a correlation of at least 0.91, with a maximum normalized RMSE and standard deviation delta of 0.58 and 0.51 respectively. To the best of our knowledge, this is the first study to investigate weather reconstruction using deep learning algorithms, proposing video inpainting as a novel approach for reconstructing missing weather information.

Plain Language Summary

We explore how deep learning can help reconstruct daily weather data. Inspired by techniques used to fill in missing parts of videos, we introduce a new method called WeRec3D, which uses a type of deep learning model that processes data in three dimensions. We improved our approach by evaluating six different techniques, resulting in a combination that is twice as accurate than our initial attempt. We further show how the location of weather observations affects the accuracy of our reconstructions and suggest a potential combination of our method with another technique from the realm of weather reconstruction. WeRec3D is trained and tested using surface temperature and pressure data over Europe. Our model achieves an average error of 1.11 °C for temperature and 199 Pa for pressure tested on the period during 1950 to 1954. As an example, we reconstruct the 1807 heat wave and validate it using a specific method that leaves out one area at a time. The results show a strong correlation with actual data and low error rates. This study is the first to use deep learning for weather reconstruction, proposing a new way to fill in missing weather data utilizing video inpainting.

1 Introduction

The recent surge in artificial intelligence (AI) has greatly intensified interest in the use of AI technologies for meteorological applications (Schultz et al., 2021). Machine learning (ML) approaches are increasingly being used to extract patterns and insights from the ever-growing stream of geodata (Reichstein et al., 2019). In this paper, we investigate the applicability of deep learning (DL) methods specifically for the task of weather reconstruction.

Extreme weather events have always occurred in the past. Researchers recorded and documented instrumental meteorological observations in analogue logbooks as early as the 17th century (Brönnimann et al., 2019; Camuffo et al., 2023). Such data is extremely valuable for climate research. Historical instrumental records of past extremes can improve our understanding of climate variability and its mechanisms. For this reason, many such logbooks have been collected, digitized, and processed in recent decades (Brugnara et al., 2020; Pfister et al., 2019). However, past meteorological observations only have limited spatial validity. In other words, they only describe the local weather in the area of the respective measuring station. This is, however, not enough to gain a better understanding of climate and to be able to run impact models which require spatially and temporally complete meteorological fields (Flückiger et al., 2017; Rössler & Brönnimann,

2018). Such detailed representations can be created from historical observations by means of weather reconstruction.

In this study, we consider weather reconstruction as the process of extending daily gridded fields of variables such as temperature or pressure backward in time. Various studies have already performed daily weather reconstructions for Europe based on the analogue resampling method, with data dating back to the 18th century (Pappert et al., 2022; Pfister et al., 2020; Imfeld et al., 2023).

Recently, methods from the field of deep learning have found their way into climate science (Gong et al., 2022; Schultz et al., 2021). For example, Kadow et al. (2020) and Yao et al. (2023) used image inpainting approaches from the realm of computer vision to reconstruct monthly climate data. Compared to monthly averages, daily observations show stronger temporal dependencies. Therefore, these cannot be modeled using an image inpainting method. Rather, in this paper we investigate the applicability of video inpainting as a new method for weather reconstruction. Video inpainting functions in a similar way to classic inpainting but operates simultaneously on several successive frames of a data stream instead of just one image at the time. Due to this additional temporal dimension, this method shares common characteristics with weather reconstruction. Both attempt to explore the spatial-temporal relationships between existing and missing data.

One of the main limitations of inpainting is that the general performance decreases as the percentage of missing parts increases (Sun et al., 2022). For historical weather reconstruction, one deals with missing rates of $\pm 99\%$, considering a spatial grid over Europe with a resolution of $1^\circ \times 1^\circ$. Thus, this is not expected to be handled accurately by an unmodified video inpainting approach. To address this challenge, we propose a novel deep learning-based weather reconstruction method, WeRec3D, based on a three-dimensional convolutional neural network. In contrast to traditional meteorological techniques that operate on anomalies (Qian et al., 2021), our method processes the climatology. It leverages an incremental pre-training approach to gradually tackle high missing rates. The use of spatially moving window sampling increases the number and variability of training examples, and thus amplifies the generalization capability. Elevation data - as a further predictor - of the corresponding areas support the orientation. To guide the learning process to a physically plausible local optimum, we apply a soft constraint to the loss function.

Within our experiments, we operate on 2-meter temperature and sea level pressure data from two periods. On the one hand, reanalysis data from the recent past (1950 to 2020) are used to train and initially validate our algorithm. These fully observed variables are artificially masked and fed to the model to learn the conditions on which the weather is based. On the other hand, we use historical weather records from the year 1807. These spatially and temporally incomplete data are reconstructed by the trained model to demonstrate its performance. The historic summer of 1807 was exceptionally warm in Europe at that time. With an anomaly of $+2.15^\circ\text{C}$, it was the warmest Alpine summer between 1500 and 1900 in the reconstruction by Casty, Wanner, et al. (2005). Due to the unusually high temperatures, this summer represents an interesting period for climate and weather research.

To the best of our knowledge, this is the first study to investigate weather reconstruction using deep learning algorithms.

2 Data

As data basis for training, we used 2-meter temperature (ta) and mean sea level pressure (slp) from the ERA5 data set (Hersbach et al., 2020). ERA5 is a global reanalysis with an hourly temporal resolution and a spatial resolution of $0.25^\circ \times 0.25^\circ$. We limited our analysis to an area of 33N to 73N and 24W to 44E and a time span from 1950 to 2020. However, the data within this period show a different statistical distribution than is the case for the inference year 1807. This manifests itself in a temperature trend

within the modern time span. Furthermore, present-day ERA5 temperatures are generally warmer than temperature observations in the early 19th century (Pappert et al., 2022; Imfeld et al., 2023). For this reason, we performed domain-specific preprocessing applied in other weather reconstruction attempts to align these data with the statistical distribution of the inference variables.

Firstly, the temporal resolution of the hourly reanalysis data was adjusted to daily mean values. Further, we reduced the spatial resolution to $1^\circ \times 1^\circ$ by average pooling. In doing so, neighbouring cells within a square are combined and replaced by a cell that corresponds to their average. This reduces the size of the matrix to be processed by a factor of 16, making computation significantly less expensive. We then ensured that the temperature data did not contain any trends in the same way as done by Imfeld et al. (2023). To do so, we calculated the zonal averages considering only land areas from the ERA5 temperature data. For each latitude, the average of the longitudes was calculated to obtain zonal averages. A linear regression model was then fitted over time to the land-only zonal mean temperature values, resulting in a slope. This slope was subtracted from the ERA5 daily temperature data for each corresponding grid cell, centered on the year 1985. To account for the higher temperatures in recent data, the climate change signal was subtracted from the training data. As in Pappert et al. (2022), we used the EKF400v2 reanalysis to determine this signal. The EKF400v2 is a global monthly climate reconstruction (Valler et al., 2021). As this covers the last 400 years, it can be used to estimate the temperature difference between the historical and modern time period. Specifically, we calculated the signal by subtracting the zonally averaged EKF400v2 temperature over land from the zonally averaged temperature of the training period over land. This resulted in a difference, which was subtracted from the processed ERA5 temperature for each latitude. Next, we divided the preprocessed data into three portions (training, validation, test) according to the block sampling presented by Schultz et al. (2021). These range from 1965 to 2020 (training), 1955 to 1964 (validation), and 1950 to 1954 (test) respectively. The first and largest block was used to train our models. The second was used to examine the quality of the individual methods comprising WeRec3D. The test set was used exclusively to assess how well our method is expected to perform on the historical observations. Since the scales of the weather variables ($^\circ\text{C}$, Pa) have a significantly different order of magnitude, we performed a z-normalization to transform them into a similar range of values.

Brönnimann (2022b) compiled a set of stations measuring temperature and pressure over Europe in 1807. The station locations were chosen so that no more than one time series occurs per $1^\circ \times 1^\circ$ grid cell. In our study, we used a collection of weather data based on this list. Table 1 lists our measuring stations and Figure 1 visualizes their position in Europe. The historical measurements show further differences with respect to the processed ERA5 grid data. On the one hand, the local conditions of the stations may have changed, on the other hand, the measuring instruments and observation methods are not the same as today. In order to account for these biases, we used homogenised data analogous to the work of Pappert et al. (2022) and Imfeld et al. (2023). To be able to process the time series using our model, we must convert them into matrix form. We defined its spatial area as 32×64 , which corresponds to a latitude and longitude of 67°N to 36°N and 22°E to 41°W . Each station position was mapped to the corresponding grid cell in which it is located. This resulted in a four-dimensional matrix $X \in \mathbb{R}^{(365, 32, 64, 2)}$. The dimensions represent days, latitude, longitude and weather variables. We scaled the values contained therein using the same statistics as for the ERA5 data.

Several studies (Gousios et al., 2023; Vaughan et al., 2022; Ivek & Vlah, 2023) have successfully used geographic information, such as topography, in climatological deep learning models. Motivated by these studies, we also investigated the effects of adding elevation data into our weather reconstructions, namely, the global elevation dataset ETOP01 (NOAA National Geophysical Data Center, 2009). To feed elevational data into our model, we needed to perform several preprocessing steps. In our target region, the data set spans

Station	ID	Variable	Lat	Lon
Armagh	ARM	slp	54.35	-6.65
Barcelona	BAR	ta	41.39	2.15
Berlin	BER	both	52.57	13.31
Cadiz	CAD	ta	36.53	-5.7
Central Belgium	CBT	ta	50.85	4.35
Central England	CET	ta	52.5	-1.9
Geneva	GVE	both	46.2	6.15
Haarlem	HAA	both	52.38	4.64
Hohenpeissenberg	HOH	both	47.8	11.02
Karlsruhe	KAR	slp	49.01	8.4
London	LON	slp	51.5	0
Milano	MIL	ta	45.47	9.19
Mulhouse	MUL	both	47.75	7.34
Padova	PAD	both	45.41	11.89
Paris	PAR	both	48.86	2.34
Prag	PRA	ta	50.07	14.41
Rovereto	ROV	both	45.9	11.05
Schaffhausen	SHA	both	47.7	8.64
Stockholm	STK	both	59.35	18.05
St. Petersburg	STP	both	59.93	30.27
Torino	TOR	both	45.07	7.68
Uppsala	UPP	ta	59.86	17.64
Valencia	VAL	both	39.48	-0.37
Vilnius	VIL	ta	54.69	25.28
Warschau	WAR	ta	52.28	20.96
Wroclaw	WRO	ta	51.11	17.03
Ylitorio	YLI	both	66.32	23.67
Zitenice	ZIT	both	50.56	14.16

Table 1. Stations used for the reconstruction of the year 1807. The abbreviations of the variables are ta - temperature; slp - pressure; both - if a station recorded the both variables.

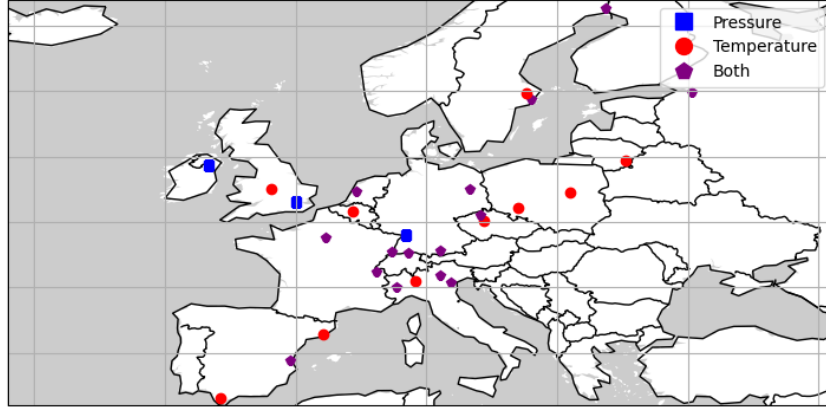


Figure 1. The spatial distribution of station locations over Europe. A blue square indicates pressure, a red circle indicates temperature, and a purple pentagon indicates both variables.

172 areas between -6374m and 5381m a.s.l. Since the ERA5 data set describes surface vari-
 173 ables, we assumed that the water depths are not relevant for modeling. Thus, we set all
 174 values below -100m to -100m. This slightly negative threshold value is intended to make

coastal regions smoother and ocean areas distinguishable from land area. We performed average-pooling on these adjusted values to achieve a spatial resolution of $1^\circ \times 1^\circ$. Finally, the values were normalized by min-max scaling.

3 Methodology

3.1 Modeling Principle

The weather reconstruction method proposed in this paper is based on video inpainting, a technique to fill spatio-temporal holes with plausible content in a video (Dahun et al., 2019). Instead of frames of a video with RGB values in the channel, we model daily temperature and pressure fields. Figure 2 illustrates the principle of the video inpainting technique for weather reconstruction and the progression of the activation maps throughout the network.

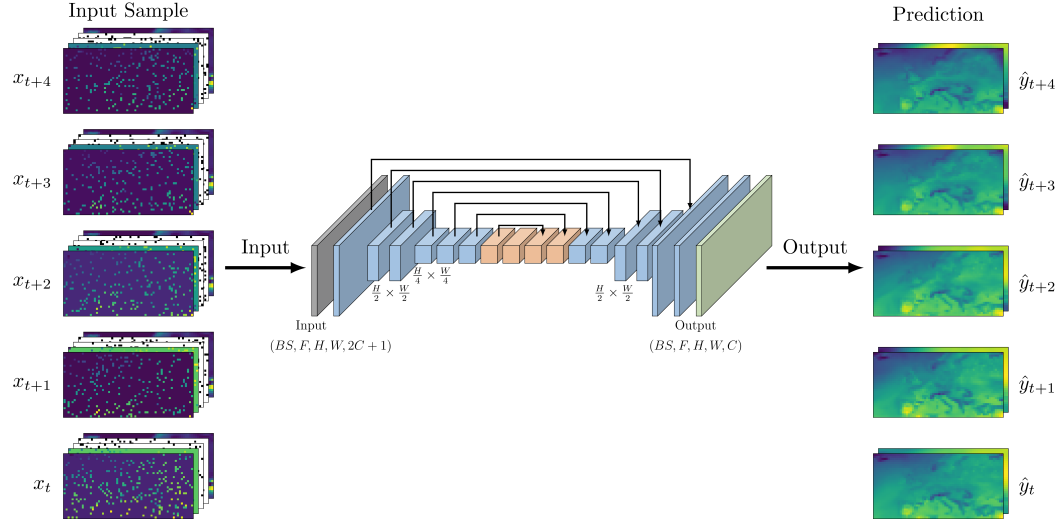


Figure 2. Modeling principle of WeRec3D. The model takes five incomplete and consecutive days as input consisting of temperature and pressure fields, binary masks and elevation data. The output is the weather reconstruction of the corresponding days.

The modeling follows a sequence-to-sequence approach. As input we use incompletely observed consecutive days. The output represents the estimate of the complete fields for the corresponding days. Furthermore, a binary mask M is contained in the input channel, which informs the model which cells to be interpreted as observed or missing. Here, 0 and 1 stand for observed and missing, respectively. For image processing, the missing pixels in the RGB data affect all channels. When reconstructing weather variables, it can be the case that air pressure was observed at a certain location but temperature was not, or vice versa. There may also be gaps in the time series. Our mask must therefore represent the presence of observations per variable and time. We therefore define our mask according to Equation 1. As the masked, i.e. missing cells, analogous to pixels, must not contain None values, they are initialized with a default number. To do so, we use the mean of the respective weather variable. This substitute value is calculated once for the training data and used in the same way for all other data sets. The last instance of the input channel represents the elevation model over Europe. Hence, WeRec3D receives an incomplete volume $X \in \mathbb{R}^{(F,H,W,2C+1)}$ and outputs the reconstruction $\hat{Y} \in \mathbb{R}^{(F,H,W,C)}$. F , H , W , C correspond to days, height, width, and channels, respectively. In our case, $F = 5$, $H = 32$, $W = 64$, and $C = 2$.

$$M \in \{0, 1\}^{(F, H, W, C)} \text{ where } \begin{cases} M_{k,i,j,c} = 1, & \text{if } V_{k,i,j,c} \text{ is missing} \\ M_{k,i,j,c} = 0, & \text{otherwise} \end{cases} \quad (1)$$

Our network is trained in a self-supervised manner. This means that the data basis for the model input and the ground truth is the same. As the training and validation data originate from a reanalysis, the meteorological fields are complete, i.e. fully observed. In order to teach the model to estimate missing cells, we need to create artificially masked fields from the complete ERA5 meteorological fields. This masking can be performed in two ways: either Missing Completely At Random (MCAR) or Not Missing At Random (NMAR). In the first case, the missing distribution is uniformly distributed across the surface. The second variant results from the actual station locations of the inference data from 1807.

3.2 Model Architecture

The WeRec3D architecture builds on the CombCN network proposed by Wang et al. (2019). In doing so, we retain the U-net-like encoder-decoder convolutional network structure. In contrast to the original two-dimensional layers, we use three-dimensional layers to seamlessly integrate the temporal dimension into a unified network. The actual number of layers and their number of filters as well as batch normalization after each layer but the last is retained. The resulting model has 22'957'736 parameters, which corresponds to a size of 90MB. Table 2 documents the details of the layers, with downward and upward arrows indicating the halving and doubling of the spatial size, respectively. In contrast to CombCN, which used the Rectified Linear Unit (ReLU) activation function, our model uses the Exponential Linear Unit (ELU). This choice is due to the centered form of our weather data, which allows for negative values unlike the RGB data of the original model, which is limited to the range $[0, 1]$. ReLU sets negative values to zero after each layer whereby ELU allows them to be propagated through the network. We, thus, assume better performance being achieved in our case by using the later activation function. Only layers 15 and 16 use the tanh activation function. They limit the corresponding activation maps to a range of -1 to +1. Finally, the identity function is used after the last layer. Using a non-limiting function at the end shall enable the network to generate potential outliers and thus effectively model extreme events. The CombCN network uses dilated convolutions in the latent space of the network to enlarge the receptive field of the output units. Thus, the kernel is inflated by artificially enlarging the spaces between the filter elements (Yu & Koltun, 2016). Since the latent space in our case is several times smaller than in the application of Wang et al. (2019), we reduce the dilation rates accordingly.

During training, our algorithm seeks to minimize the loss function (Equation 4), which consists of a linear combination of a masked mean absolute error (MAE) (Equation 2) and a normal MAE (Equation 3). Where \odot is the pixelwise multiplication, $\|\cdots\|$ is the l_1 -norm, Y is the ground truth, and BS is the batch size. α controls the composition of the linear combination and is set to 0.5. In this way, WeRec3D is taught to generate the entire meteorological field with a stronger focus on the masked cells. This is because in Equation 2, the errors in the reconstruction of the originally observed cells are multiplied by zero and therefore not taken into account.

3.3 Physical Soft Constraint by means of Covariance Matrix Inclusion

Off-the-shelf deep neural networks do not necessarily obey the fundamental laws of physical systems (Kashinath et al., 2021). As a result, model outputs can potentially become physically impossible. However, it is vital to ensure such plausibility, especially when making predictions for situations for which the model has not been explicitly trained. Regularization techniques are used in machine learning to prevent overfitting, for exam-

Layer No.	Type	Kernel	Stride	Filters	Dilation	Padding	Activation
1	conv.	(3, 5, 5)	(1, 1, 1)	64	-	same	ELU
2	conv. ↓	(3, 4, 4)	(1, 2, 2)	128	-	valid	ELU
3	conv.	(3, 3, 3)	(1, 1, 1)	128	-	same	ELU
4	conv. ↓	(3, 3, 3)	(1, 2, 2)	256	-	valid	ELU
5	conv.	(3, 3, 3)	(1, 1, 1)	256	-	same	ELU
6	conv.	(3, 3, 3)	(1, 1, 1)	256	-	same	ELU
7	dilated conv.	(3, 3, 3)	(1, 1, 1)	256	(1, 2, 2)	same	ELU
8	dilated conv.	(3, 3, 3)	(1, 1, 1)	256	(1, 2, 2)	same	ELU
9	dilated conv.	(3, 3, 3)	(1, 1, 1)	256	(1, 3, 3)	same	ELU
10	dilated conv.	(3, 3, 3)	(1, 1, 1)	256	(1, 4, 4)	same	ELU
11	conv.	(3, 3, 3)	(1, 1, 1)	256	-	same	ELU
12	conv.	(3, 3, 3)	(1, 1, 1)	256	-	same	ELU
13	deconv. ↑	(1, 3, 3)	(1, 2, 2)	128	-	valid	ELU
14	conv.	(3, 3, 3)	(1, 1, 1)	128	-	same	ELU
15	deconv. ↑	(1, 3, 3)	(1, 2, 2)	64	-	valid	tanh
16	conv.	(3, 3, 3)	(1, 1, 1)	32	-	same	tanh
17	conv.	(3, 3, 3)	(1, 1, 1)	2	-	same	linear

Table 2. Network architecture of the WeRec3D model. In the type column, ↓ and ↑ indicate the halving and doubling of the spatial size, respectively.

$$MMAE(Y, \hat{Y}, M) = \frac{1}{BS} \frac{1}{F} \sum_{b=1}^{BS} \sum_{k=1}^F \frac{\|M_b^k \odot (Y_b^k - \hat{Y}_b^k)\|}{\|M_b^k\|} \quad (2)$$

$$MAE(Y, \hat{Y}) = \frac{\|Y - \hat{Y}\|}{BS \cdot F \cdot H \cdot W \cdot CH} \quad (3)$$

$$\mathcal{L}(Y, \hat{Y}, M) = \alpha \cdot MMAE + (1 - \alpha) \cdot MAE \quad (4)$$

ple, by extending the optimization function. Similarly, physical information can be incorporated by enhancing the loss function with prior knowledge. For instance, additional terms can be included that describe physical relationships in the mappings. This transforms the optimization space and can promote the convergence of the training process to more plausible solutions (Jia et al., 2021). Such regularization only provides some guidance towards a physically sound optimum, but does not enforce it and is thus also referred to as a soft constraint (Kashinath et al., 2021).

In weather reconstruction, field pattern analysis provides information about the relationship between different variables and areas in space and time. For this purpose, climate researchers use a principal component analysis, which is derived from the covariance matrix of the meteorological fields (Luterbacher et al., 2002; Casty, Handorf, et al., 2005). The covariance matrix describes the relationship between each grid cell and every other position in the area under consideration. For example, Iceland and the Azores have an inverse relationship with regard to air pressure (Stephenson et al., 2003). If the pressure rises over Iceland, it falls over the Azores and vice versa. Such correlations can also be seen for the temperature and intravariation in the covariance matrix.

Our application should be able to correctly reproduce such relationships in the reconstruction. Therefore, we develop a physical soft constraint that informs the model of potentially misaligned correlations between the predictions cells. The soft constraint is implemented via an extension of the loss function according to Equation 5. Specifically, it is the MAE of two covariance matrices, i.e. $\ell(\sigma_y^2, \sigma_{\hat{y}}^2) = \frac{1}{n} \sum_{i=0}^n |\sigma_{y_i}^2 - \sigma_{\hat{y}_i}^2|$. With $\sigma_{\hat{y}}^2$ being calculated on the prediction of the model, σ_y^2 on the corresponding ground truth. The process of creating a covariance matrix is the same for both and is visualized in Figure 3. In the course of each training iteration, WeRec3D processes a batch of random samples. Each sample contains five consecutive days, i.e. frames in the form of temperature and air pressure fields. First, we flatten the meteorological fields and concatenate their variables, creating a one-dimensional row vector for each day. The vectors are then stacked to obtain a 2D matrix of the form days \times cells. We then calculate the covari-

ance matrix of this using the equation shown on the right of the figure. The formula is chosen in matrix form to enable a high-performance implementation. Comprising continuous algebraic operations, it is inherently differentiable, which is a key ingredient for backpropagation.

The constraint is added to the objective function through a linear combination controlled by a hyperparameter β . We set this to 0.9 so that the proportion of the physical soft constraint is 10%. In this way, the constraint slightly guides the learning process while the original loss part remains decisive.

$$\mathcal{L}(Y, \hat{Y}, M) = \beta \cdot (\alpha \cdot \text{MMAE} + (1 - \alpha) \cdot \text{MAE}) + (1 - \beta) \cdot \ell(\sigma_y^2, \sigma_{\hat{y}}^2) \quad (5)$$

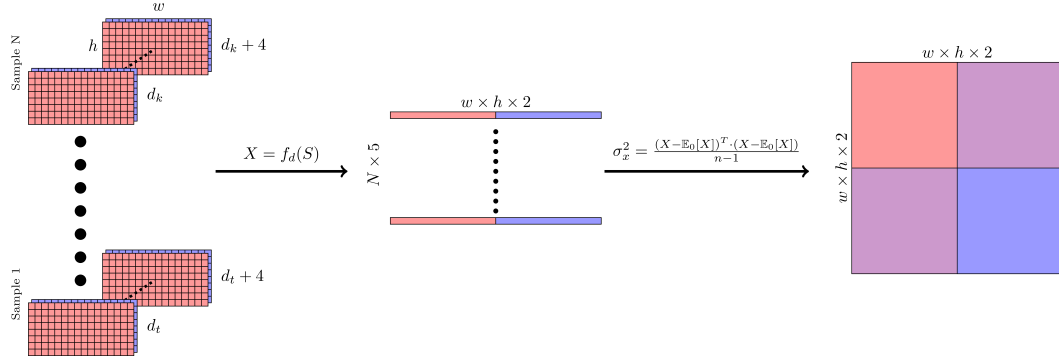


Figure 3. Covariance Matrix Creation. Temperature and pressure are represented by the colors red and blue, respectively. In the matrix on the right, these colors correspond to their respective covariance values. The purple squares indicate the covariance values between temperature and pressure.

3.4 Training Details

The performance of inpainting tasks decreases with increasing missing rate. Therefore, we expect that our model will have severe difficulties in learning and reconstructing weather phenomena based only on 1% observed cells. To overcome this obstacle, we apply a training method, which we call incremental pre-training. Its aim is to gradually familiarize the model with the actual reconstruction task by training it successively on 10%, 20%, ..., 90% and 99% missing rates. At each level, we use 10 epochs. Instead of reinitializing the parameters of our network for the next percentage level, we use the weights of the previous trained model. In other words, each reconstruction can benefit from what has already been learned and build on it. In particular, the use of the normal MAE loss component should contribute significantly to this, because it enables WeRec3D to learn from the non-masked cells as well. We assume that relevant weather features will be learned at low percentage levels and that this knowledge will be transferred up to the 99% rate. However, this approach also poses a risk. Processing the same examples multiple times could lead to the memorization of the training samples. To minimize overfitting, we only pass the weights resulting from early stopping to the next initialization. That is, from the epoch that results in the lowest validation error.

Generally, the more data a deep learning algorithm receives for training, the better it performs. In the field of computer vision, the amount of training data is often artificially increased by data augmentation. Data augmentation creates new samples from existing images or frames by cropping, rotating, distorting, or shading them with a color tone (Shorten & Khoshgoftaar, 2019). These manipulations also increase the variability of the

data, which reduces overfitting of models. The way in which the examples can be changed depends always on the nature of the problem. In other words, the results of the manipulation must still make sense in the context of modeling. For our use case, this means that external influences on European weather, such as jet streams, must remain valid. We therefore consider augmentations such as rotation or mirroring to be impractical. However, changing the spatial position seems sensible. The blue rectangle in Figure 4 shows the base window of size 32x64. Our data augmentation approach now consists of moving this window in order to generate slightly different samples. Specifically, during training, each batch of samples is randomly manipulated twice. Thereby, the window size remains the same, but the position comes to lie in the expanded region (33N to 73N and 24W to 44E). The potential shifts are illustrated by the red rectangles. For the validation of the reconstruction, we only use the base window. WeRec3D uses Glorot initialization (Glorot & Bengio, 2010) for weights, minimizes the loss function with the default parameterised Adam optimization (Kingma & Ba, 2017) during training, and processes data in batches of 16 samples.

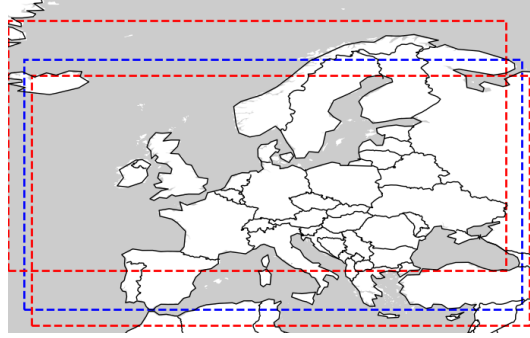


Figure 4. Illustration of Moving Window Sampling. The default or base window is indicated by the blue rectangle, possible shifted areas are indicated by the two red rectangles.

3.5 Creating WeRec3D: Building the Leading Method through Iterative Enhancements

WeRec3D leverages various strategies to enable an accurate reconstruction capability when rates of missing data are high. In this section, we outline the iterative evaluation process that led to our final methodology. Specifically, this involves six enhancement techniques that were included or omitted in the training process. **(i) Use of climatology (Clim.)** instead of the anomaly used in classical weather analysis. **(ii) Incremental pretraining (IPT)** instead of directly modeling a missing rate of 99%. **(iii) Spatially moving window (MovWin):** During training, we feed the network with meteorological fields that are sampled at different spatial locations rather than just at a fixed position. **(iv) Soft constraint on the loss function (SC):** A physical restriction that steers the learning process towards a more plausible local optimum. **(v) Elevation data (Elev.)** as a further predictor. **(vi) Physically-informed initialization (PII):** This technique is borrowed from Yao et al. (2023) and aims to start the inpainting process with a relatively reliable and physically plausible basis in the missing areas. Instead of using the same initial value for all masked cells, each cell is individually initialized with the corresponding average value over time.

It is impossible to predict in advance whether the strategies introduced and their combinations will positively or negatively affect the quality of the reconstruction. Analyzing all possible combinations would yield 64 method variants, as $2^6 = 64$. Therefore, our objective is to streamline the search for the most effective combination. To achieve this, we apply an iterative heuristic to efficiently find a leading method in its local neigh-

borhood. Thereby, we imagine the combination of techniques as a tree. The root node is the baseline model, followed by all six strategies on the first level. Each node is succeeded by the techniques that were not previously encountered on the path leading to the node. Our iterative discovery procedure proceeds as shown in the enumeration below. Depending on the improvement of an additional extension, the tree is traversed in a depth-first-search or a breadth-first-search manner.

1. We start by naming the base model, which serves as the first reference point, the *leading method*. We then move on to the first level of the tree.
2. Select a node at the current tree level that has not yet been visited.
 - (a) We analyze the resulting accuracy and compare it to that of the *leading method*.
 - (b) If the performance is significantly better, this approach becomes the new *leading method*. We then traverse the tree one level further and start again at Step 2.
 - (c) If the quality is not significantly better, we visit the next unvisited node at the current level and repeat Step 2.a).
 - (d) If no significantly better performance could be achieved, we choose the technique that most reduced the error. This is now considered the *leading method*. We go down one level along the corresponding node and move back to Step 2.
 - (e) If no technique was able to reduce the error, the search is complete. The same applies if we have reached a leaf node.
3. After completion of the search, the combination marked as *leading method* is our choice for the best modeling, that is, WeRec3D.

We set the significance level to 90%. Consequently, a newly added technique must reduce the error compared to the last leading method to at least 90% to be considered a significant improvement. In addition, we set the random seed to the same value for each experiment to make them comparable to each other. Thus, the weights are initialized in the same way each time, and the shuffle of the samples results in the same order for each trial. In addition, the same masking, i.e. the artificial deletion of cells, is used in every experiment. The cells to be interpreted as observed or missing are based on a random uniform distribution over the field.

Section 4.1 shows the results of this evaluation. The best validation metrics were achieved by combining the techniques **i** to **v**.

3.6 Accounting for the Station Distribution in 1807

Different historical periods may exhibit a different spatial distribution of the weather stations at that time. For the inference of the year 1807, this corresponds to the positions in Figure 1. Consequently, all fields have the same observed cell positions, apart from the few gaps in the temporal dimension. Every cell in which there is no station is to be regarded as missing. We refer to such masking as *Not Missing At Random (NMAR)*. For the evaluation described in Section 3.5, however, we used a *Missing Completely At Random (MCAR)* distribution of the observations. This means that the positions of the missing cells are set randomly in the spatial dimension. Thus, in contrast to NMAR, all fields have different observation positions. We use MCAR masks to assess the techniques independently of a specific historical event and its weather station positions. However, we assume that a reconstruction based on an NMAR input is more difficult than on MCAR observations, because the variability of the information available to the model is reduced. To evaluate this assumption, we analyzed how the resulted leading method performs when given NMAR-masked validation data as input. Next, we compared the corresponding accuracy with two approaches that specifically adjust the model to the NMAR distribution: *NMAR training* and *NMAR fine-tuning*. This comparison was still

made based on the validation set. Only the best-performing strategy will be further examined based on the test set in Section 4.3 to provide an estimate for inference. With *NMAR training*, we incrementally train at several percentage levels in the sense of incremental pre-training. However, the observations are now statically defined as the same for all samples per variable. Figure 5 illustrates the corresponding masks of the temperature variable. The white cells indicate masked areas, and the black pixels represent the locations treated as observed. Thereby, the observed region decreases spherically in the direction of the station positions as the missing rate increases. The 99% level thus corresponds to the observation locations of the target year 1807. For the second approach, *NMAR fine-tuning*, the model is not taught from scratch. Instead, we fine-tune the leading model which was trained on a MCAR distribution to focus specifically on the observation positions of 1807.

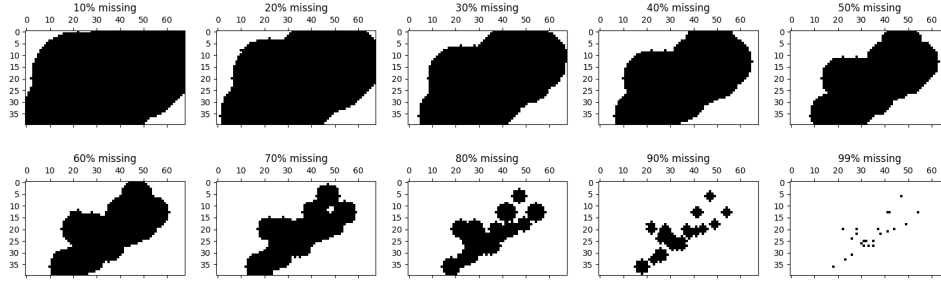


Figure 5. Masks used for NMAR training. White and black represent the missing and observed areas, respectively. The region of observed cells decreases spherically towards the station positions of 1807.

3.7 Artificially Increasing the Proportion of Observations Through ARM Enhancement

Our analyzes showed that the validation error increases exponentially between 90% and 99% missing rates. Therefore, the reconstruction accuracy for the year 1807 could presumably be significantly improved if a small number of additional observations were available. However, the retrieval of further historical weather logbooks is very time-consuming. For this reason, we are investigating whether the reconstruction quality can be improved by artificially increasing the proportion of observations. By an artificial increase we mean the input of additional cell values that are plausible in relation to the respective field per day. Specifically, we borrow these values from the Analogue Resampling Method (ARM) (Lorenz, 1969).

ARM is a common and successful statistical model used for weather reconstruction (Brönnimann, 2022a; Pappert et al., 2022; Pfister et al., 2020; Imfeld et al., 2023). The model assumes atmospheric state patterns will repeat themselves over time. If one state is similar to another, the resulting local weather effects are presumably similar (Lorenz, 1969). The idea of ARM-based weather reconstruction is in principle comparable to the kNN algorithm with $k = 1$. One takes a historical day whose field was not completely observed, i.e. a day whose missing field cells one wants to reconstruct. From a pool of fully observed fields, the ARM selects the day that best matches the observation locations of the historical day. Based on the above assumption, the two days have similar local weather effects. The selected day is called the best analogue. It represents the reconstruction of the weather in the unobserved areas in the past (Brönnimann, 2022a).

For each incomplete day in 1807, we searched for a similar complete field from the ERA5 training dataset. We then randomly replaced 3% of the unobserved cells of each historical field with the cell values of the respective best analogue and treated them as observed.

This artificially decreases the missing rate of the inference data from 99% to 96%, resulting in observations that approximate a MCAR distribution. In accordance with this distribution, we refined the training process of WeRec3D to reconstruct the year 1807 with this enhanced input. We have again trained the model on several MCAR percentage levels, but replaced the final 99% mask with a 96% mask. The corresponding results are shown in Section 4.4.

4 Results and Discussion

4.1 Evaluation of the Creation of WeRec3D.

In this section, we analyse the iterative creation of WeRec3D. Table 3 shows the experiments performed and their performance on the validation data at 99% MCAR. For better comparability, we calculate all validation metrics based on the anomalies of the prediction and the ground truth. The order from top to bottom corresponds to the discovery procedure described in Section 3.5. The six columns on the right-hand side show which extensions were included in the experiment. In addition to the specific validation MAE value, we indicate whether the experiment led to a leading method (LM) and to which degree the error has decreased (Err. Dec.) compared to the previous LM. Our finding procedure starts with evaluating the baseline (1). This means only training the convolutional network on anomalies and without additional extensions. The first evaluated strategy, the preservation of climatology (2), was already able to significantly reduce the validation error to 77%. Therefore, we kept this approach directly and investigated incremental pre-training (3) based on it. This further reduced the error of the current leading method to 67%. The next four experiments (3.1 to 3.4) operate at the same tree level. At this point, no technique was able to significantly improve the quality. Only the addition of elevation data as a further predictor has a minimal positive effect. For this reason, the corresponding node becomes the new leading method. Three extensions remain, which are validated in experiments 4.1 to 4.3. The spatially moving window method is the only one that slightly reduces the error. After its contribution, two potential approaches remain (5.1 and 5.2). At this level, the physical soft constraint (5.1) wins the race, although not significantly. After that, the physically informed initialization (6.1) fails in providing better performance. A noticeable finding in this analysis is the synergy between *Elev*, *MovWin* and *SC*. Initially, neither *MovWin* nor *SC* was able to reduce the error. Only when combined with the additional elevation data did the moving window technique lead to a reduction in the validation MAE. We assume that without the topography, the model has difficulties distinguishing between the differently sampled window positions. These difficulties seem to outweigh the added value of data augmentation through *MovWin*. However, as soon as the orientation of the model is supported by the topological input, the reconstruction quality improves. The situation is similar with the physical soft constraint (*SC*). This improvement technique only has a positive influence when it is combined with *MovWin*. The soft constraint is based on the covariance matrix, which is calculated using the fed fields. If the position of the fields remains the same and there are enough samples (in our case $N = 80$), the covariance matrix should always appear fairly consistent. In this case, the model can only gain limited information from it. The *MovWin* technique changes the field position for each batch. As a result, the information content about the meteorological correlations also increases, which argumentatively leads to an increase in accuracy.

The heuristically best combination (marked with *) was able to reduce the error compared to the baseline to 48%. Scaled back to the actual units, this results in a MAE of 0.85 °C and 122 Pa for temperature and pressure respectively. The spatial and temporal distribution of the MCAR validation error can be seen in Figures 6 and 7. Our model exhibits the most pronounced challenges in accurately reconstructing temperatures within the regions of Iceland and Norway. In these areas, the average error is greater than 2 °C.

Experiment	Val MAE	LM	Err. Dec.	Clim.	IPT	MovWin	SC	Elev.	PII
1: Baseline	0.2144	✓	-						
2: Clim.	0.1654	✓	77%	✓					
3: IPT	0.11064	✓	67%	✓	✓				
3.1: MovWin	0.1244	-	-	✓	✓	✓			
3.2: SC	0.1138	-	-	✓	✓		✓		
3.3: Elev.	0.1106	✓	99.96%	✓	✓			✓	
3.4: PII	0.1131	-	-	✓	✓				✓
4.1: MovWin	0.1064	✓	96%	✓	✓	✓		✓	
4.2: SC	0.1133	-	-	✓	✓		✓	✓	
4.3: PII	0.1142	-	-	✓	✓			✓	✓
5.1: SC (*)	0.1032	✓	97%	✓	✓	✓	✓	✓	
5.2: PII	0.1084	-	-	✓	✓	✓		✓	✓
6.1: PII	0.1093	-	-	✓	✓	✓	✓	✓	✓

Table 3. Validation errors of differently combined enhancement techniques, with * marking the lowest error. In the LM column, a ✓ indicates whether a technique led to a leading method. Err. Dec. shows the amount by which the error was reduced compared to the previous leading method. A ✓ in the remaining cells indicates the investigated techniques in each experiment (see Section 3.5 for abbreviations).

In Central Europe, the temperature error is around 1°C; over the sea it is even less than 0.5 °C. The situation is different for pressure. Here, only the areas at the edge of our window cause difficulties; a phenomenon that is typical for CNN-based models. The spatially summarized errors show a clear seasonality in both variables. The winter months appear to be reconstructed with less accuracy and higher variability than the summer months. Particularly noteworthy is the outlier of the temperature error in February, which is almost 3 °C. According to Portenier et al. (2017), Western Europe was hit by a severe cold wave in February 1956, which led to exceptionally low temperatures in the region. Our model has great difficulty in accurately reconstructing the conditions at that time. In fact, the five highest temperature errors on the validation set occur between 1956.01.30 and 1956.02.05 with $\mu = 2.05$ °C and $\sigma = 0.34$ °C.

4.2 Evaluation based on NMAR Distributed Observations

In this section, we investigate how the leading method performs on NMAR input and compare it to the alternative strategies *NMAR training* and *NMAR fine-tuning*. Table 4 presents the validation metrics for the leading method using MCAR and NMAR input, as well as for the alternative strategies using NMAR input. The performance of the leading method approximately halves on NMAR input compared to MCAR input. The alternative training strategies, which are designed for the specific station distribution of 1807, show significantly improved performance under NMAR conditions. The two outcomes are nearly indistinguishable, both showing practically the same normalized MAE values. Only when the metrics are scaled proportionally to the share of each variable do subtle differences emerge. Given that we place greater importance on improvements in the temperature variable compared to the pressure variable, we have opted for NMAR training as our training strategy for subsequent inference tasks. However, this approach still has a significantly higher error than if the input were MCAR. This is justified by the fact that with NMAR, the model must extrapolate into regions where it has no prior knowledge, whereas with MCAR, it predominantly interpolates between observed data

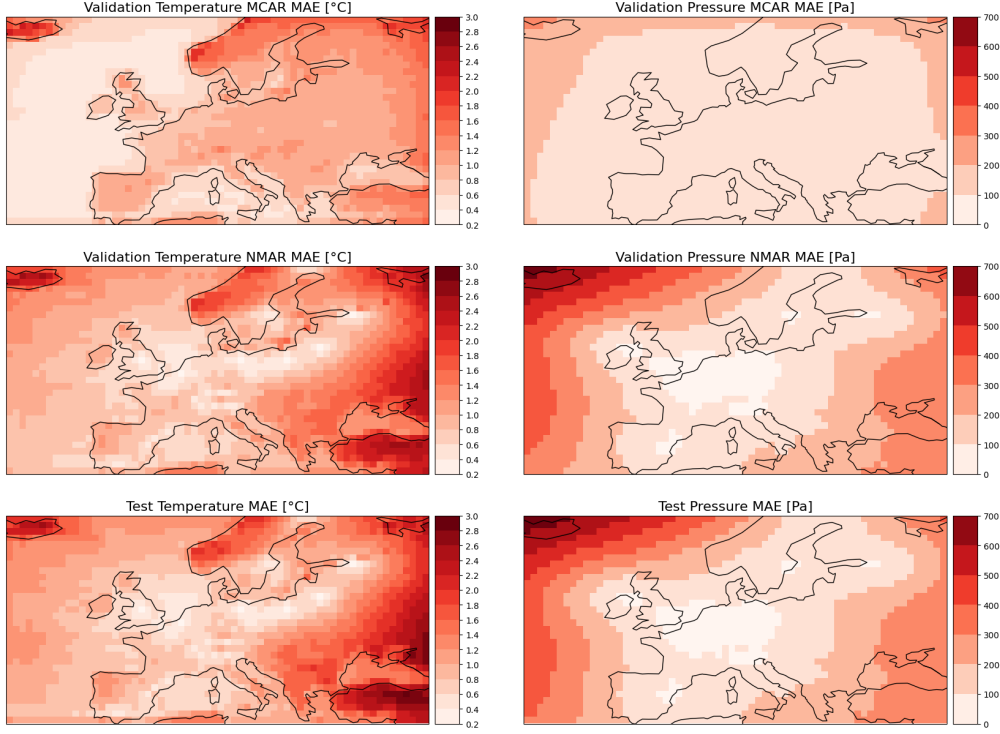


Figure 6. Spatial error of temperature and pressure over time during 1955 to 1964 for validation and during 1950 to 1954 for test. The top row results from MCAR validation data modeled by the leading method, the middle row from NMAR validation data modeled by the NMAR trained model, and the bottom row from the test data modeled by the NMAR trained model.

points. Figures 6 and 7 (middle row) display the spatial and temporal validation error resulting from the NMAR training strategy. The NMAR distribution of the observations has a strong effect on spatial quality. In Central Europe, the area with a high density of observations, errors tend to be low. However, for both temperature and pressure, errors increase with distance from the areas with a high spatial coverage of observations. The error over time exhibits seasonality with larger errors in the winter months compared to the summer months.

Strategy	Input distribution	Val MAE	ta [°C]	slp [Pa]
Leading Method	MCAR	0.103	0.85	122
Leading Method	NMAR	0.238	1.8	301
NMAR Fine-tuning	NMAR	0.154	1.15	196
NMAR training	NMAR	0.153	1.11	199

Table 4. Validation errors on Completely-Missing-At-Random and Not-Missing-At-Random inputs with 99% missing rate.

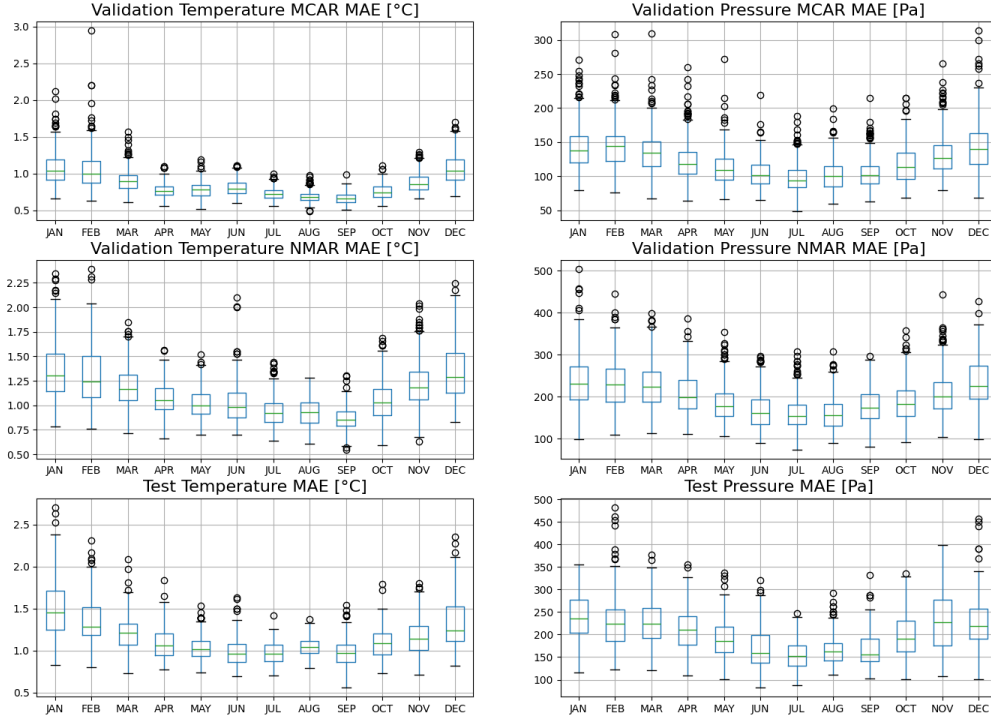


Figure 7. Temporal error of temperature and pressure over time during 1955 to 1964 for validation and during 1950 to 1954 for test. The top row results from MCAR validation data modeled by the leading method, the middle row from NMAR validation data modeled by the NMAR trained model, and the bottom row from the test data modeled by the NMAR trained model.

4.3 Evaluation based on the Test Set

We have now arrived at an appropriate methodology for inference, i.e. the reconstruction of the weather in 1807, by training WeRec3D from scratch using NMAR masks as shown in Figure 5. Up to this point, all evaluations of our approaches have been based on the validation data set. Thus, one can expect that the chosen method exhibits an artificial skill to a certain extent. To more accurately assess the quality of the model, we are conducting a weather reconstruction on the test set (1950 to 1954) in this section. The test set anomaly MAE amounts to 0.156 (1.15 °C and 201 Pa) given a NMAR input with 99% missing cells. This error represents a mere 2% increase compared to the validation data, suggesting a robust generalization capability of our methodology. Figures 6 and 7 (lowest row) display its spatially and temporally distributed errors, which strongly resemble the errors of the NMAR validation data (middle row).

4.4 Evaluation of the Reconstruction over Europe in 1807

In this section, we perform the actual reconstruction of the historical weather measurement of 1807. In contrast to the previously used reanalysis data, we do not have complete ground-truth fields available for validation. Since we only have 43 time series of the measuring stations for the year 1807, we cannot do a full spatial assessment. Instead, we carry out a leave-one-out (LOO) procedure over space by running 43 reconstructions and for each run leaving out one time series. The thereby generated predictions of the corresponding cell are then compared with the omitted observation. Some processing steps

are required to perform the LOO cross-validation. First, the seasonality in the temperature data has to be removed since otherwise, the correlation is influenced by the annual cycle of temperature. Therefore, we fit the first two harmonics using linear regression as it has been done by Pfister et al. (2020). This results in a sine-like function that is subtracted from the corresponding temperature data. For pressure data, we validate on its daily anomaly, calculated with respect to the deviation from the long-term average of each day of the year. To display all measurements in the same plot, they are normalized by dividing the RMSE and the two standard deviations by the standard deviation of the corresponding observation. All observations have, therefore, the same reference point and show uniform metrics for their prediction. In the following, we begin with the evaluation of the weather reconstruction using the WeRec3D model trained on NMAR masks. Then we show the performance of the reconstruction whose input was enhanced with ARM cells as described in Section 3.7.

Figure 8 (top) shows the quantitative metrics resulting from the LOO procedure for the 43 time series in a Taylor plot (Taylor, 2001). About half of the values are closely centered around the optimal reference point, indicating a high accuracy of the reconstruction. Each time series was reconstructed with at least a correlation of 0.91 and a maximum normalized RMSE and standard deviation delta of 0.58 and 0.51 respectively. The best (St. Petersburg) and worst (Central Belgium) reconstructed temperature time series are shown in Figure 9. This comparison is intended to illustrate the range of quality of the remaining reconstructions. St. Petersburg has the lowest and the Central Belgium time series the highest normalized RMSE value among the anomaly temperatures. In the Taylor plot, the Belgian station corresponds to the blue marker at the top right. The most accurate prediction is almost congruent with the target series. Similar qualities can therefore be expected for markings close to the reference point. The worst prediction (Central Belgium) tends to have a negative bias in winter and a positive bias in summer. Depending on the day, the absolute error can be greater than 5 °C. However, the normalized correlation with the historical observation is still over 0.95, which can be clearly seen in the graph. This is because even if the prediction often overshoots the target, the deflection of the curve heads nevertheless generally in the right direction.

The results produced using ARM enhanced input differ from the variant without it in a subtle but potential important characteristic. As can be seen in Figure 8 (bottom), there is now a small gap between the reference point and the best reconstruction. This means an increase in the normalized RMSE and a very slight degradation of the correlation for the best reconstructions. However, it can be clearly seen that the normalized standard deviations of the predictions are now less broadly distributed and are increasingly on the left side ($\sigma < 1$). The variance of these predictions, thus, tends to be lower than that of the prediction without ARM enhancement and also lower compared to the corresponding reference series. The reconstructed stations show a correlation of at least 0.84 and a maximum normalized RMSE and standard deviation delta of 0.54 and 0.27 respectively.

5 Conclusion

In this paper, we set out to investigate the potential of artificial intelligence for weather reconstruction. As a result, we propose a tailor-made network architecture, called WeRec3D, which has been optimized by innovative extensions of the modeling process to the extrapolation of daily pressure and temperature fields. The resulting method allows the reconstruction of historical weather observations describing only one percent of the area in Europe on a $1^\circ \times 1^\circ$ resolution grid, i.e. a gap filling of data with 99% missing rate. Furthermore, the modular design of our solution allows the inclusion of additional weather variables as well as the use of different resolutions of the space-time volume. This makes it suitable for reconstructing arbitrary historical events.

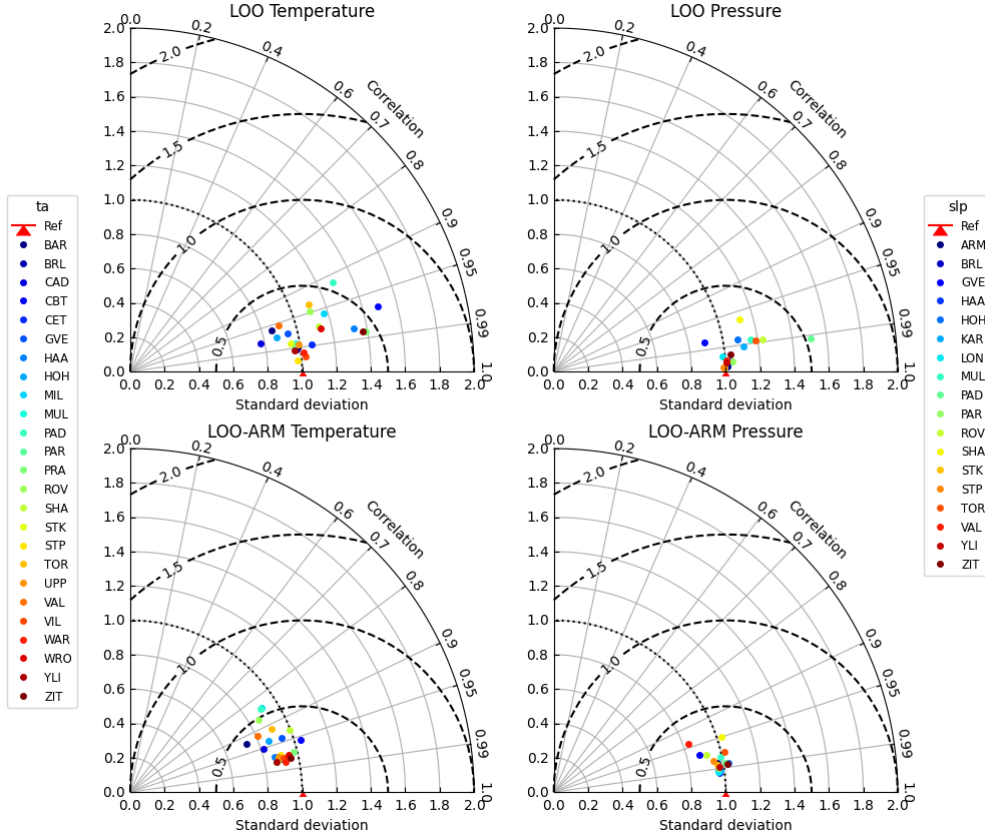


Figure 8. Leave-one-out validations in space of 1807 using WeRec3D trained on NMAR masks (top) and using ARM enhanced input (bottom).

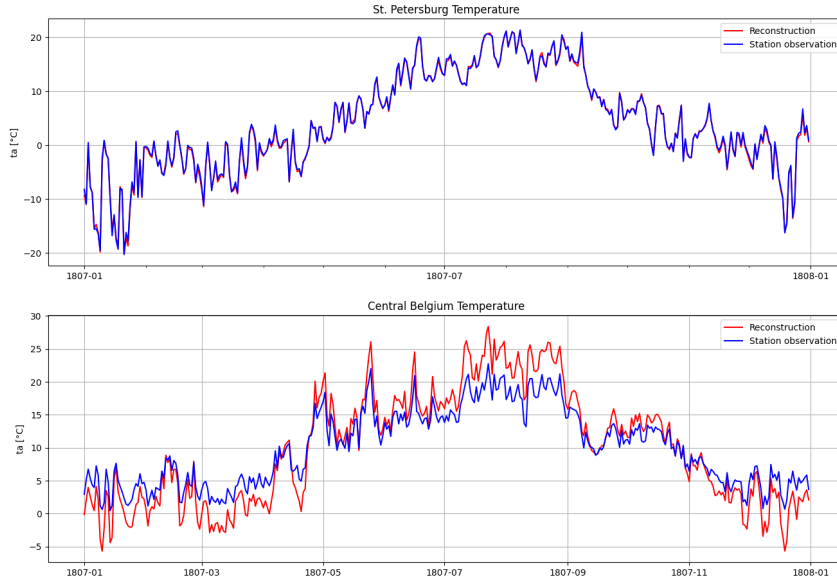


Figure 9. The temperature reconstructions of 1807 yielding the best (top) and worst (bottom) RMSE score. The ground truth and reconstruction are shown in blue and red, respectively.

The basic building block of our method is a neural network consisting of three-dimensional convolutional layers. These enable the simultaneous modeling of space and time. The adopted modeling extensions in combination reduce the reconstruction error of WeRec3D by factor two. In contrast to classical weather analysis, our method works best on pure climatology and not on its anomaly. The use of a moving-window method to sample the weather fields leads to an increased variability of the inputs and thus to an amplified generalization capability. Elevation data - as a further predictor - of the corresponding areas support the orientation. To guide the learning process to a physically plausible local optimum, we apply a soft constraint to the loss function. This is derived from the covariance matrix of the meteorological fields, which describes the intervariable relationships between temperature and pressure. Through incremental pre-training on successively increasing error rates, the model learns weather patterns. The acquired knowledge can then be reproduced even if the input is 99

The type of distribution of the observation positions has a decisive influence on the quality of the reconstruction, meaning that the accuracy of randomly distributed measurements is approximately twice as high as if they are distributed at fixed positions. However, the performance of weather modeling can be considerably improved if the algorithm is specifically trained on the positions of the expected observations. Alternatively, the artificial reduction of the missing rate using the analogue resampling method offers a promising solution. To verify the effectiveness of our reconstruction model, validation was performed on both recent and historical data. The analysis of the year 1807 shows a strong correlation and marginal RMSE values during the LOO validation in space across the weather stations.

Future work will explore the application of WeRec3D to other regions and periods, the improvement of its interpretability and the incorporation of further predictors.

Open Research Section

The code and data required to replicate the results discussed in this paper are available on GitHub and have been archived with a DOI. You can find the code at <https://github.com/YannisSchmutz/WeRec3D/tree/v1.0.0> (Schmutz, 2024).

Acknowledgments

SB and NI acknowledge funding from the Swiss National Science Foundation (grant no. 188701).

References

- Brönnimann, S. (2022a, 06). From climate to weather reconstructions. *PLOS Climate*, 1, e0000034. doi: 10.1371/journal.pclm.0000034
- Brönnimann, S. (2022b, 04). Historical observations for improving reanalyses. *Frontiers in Climate*, 4, 880473. doi: 10.3389/fclim.2022.880473
- Brugnara, Y., Pfister, L., Villiger, L., Rohr, C., Isotta, F. A., & Brönnimann, S. (2020). Early instrumental meteorological observations in Switzerland: 1708–1873. *Earth System Science Data*, 12(2), 1179–1190. doi: 10.5194/essd-12-1179-2020
- Brönnimann, S., Allan, R., Ashcroft, L., Baer, S., Barriendos, M., Brázdil, R., ... Wyszniński, P. (2019). Unlocking pre-1850 instrumental meteorological records: A global inventory. *Bulletin of the American Meteorological Society*, 100(12), ES389 - ES413. doi: 10.1175/BAMS-D-19-0040.1
- Camuffo, D., della Valle, A., & Becherini, F. (2023). Instrumental and observational problems of the earliest temperature records in Italy: A methodology for data recovery and correction. *Climate*, 11(9). Retrieved from <https://www.mdpi.com/2225-1154/11/9/178> doi: 10.3390/cli11090178

- 637 Casty, C., Handorf, D., Raible, C., González-Rouco, J., Weisheimer, A., Xoplaki, E.,
638 ... Wanner, H. (2005). Recurrent climate winter regimes in reconstructed and
639 modelled 500 hpa geopotential height fields over the north atlantic/european
640 sector 1659–1990. *Climate Dynamics*, 24, 809–822.
- 641 Casty, C., Wanner, H., Luterbacher, J., Esper, J., & Böhm, R. (2005, 11). Tempera-
642 ture and precipitation variability in the european alps since 1500. *International*
643 *Journal of Climatology*, 25, 1855–1880. doi: 10.1002/joc.1216
- 644 Dahun, K., Woo, S., Lee, J.-Y., & Kweon, I. (2019, 06). Deep video inpainting. In
645 (p. 5785–5794). doi: 10.1109/CVPR.2019.00594
- 646 Flückiger, S., Brönnimann, S., Holzkämper, A., Fuhrer, J., Krämer, D., Pfister, C.,
647 & Rohr, C. (2017). Simulating crop yield losses in switzerland for historical
648 and present tambora climate scenarios. *Environmental Research Letters*, 12(7),
649 074026. doi: 10.1088/1748-9326/aa7246
- 650 Glorot, X., & Bengio, Y. (2010, 13–15 May). Understanding the difficulty of
651 training deep feedforward neural networks. In Y. W. Teh & M. Titterton
652 (Eds.), *Proceedings of the thirteenth international conference on artificial*
653 *intelligence and statistics* (Vol. 9, pp. 249–256). Chia Laguna Resort, Sar-
654 dinia, Italy: PMLR. Retrieved from [https://proceedings.mlr.press/v9/](https://proceedings.mlr.press/v9/glorot10a.html)
655 [glorot10a.html](https://proceedings.mlr.press/v9/glorot10a.html)
- 656 Gong, B., Langguth, M., Ji, Y., Mozaffari, A., Stadtler, S., Mache, R. K., & Casty,
657 M. (2022, 12). Temperature forecasting by deep learning methods. *Geoscientific*
658 *Model Development*, 15, 8931–8956. doi: 10.5194/gmd-15-8931-2022
- 659 Gousios, G., Mamouka, T., Vourlioti, P., & Kotsopoulos, S. (2023, 06). *Downscal-*
660 *ing seasonal weather forecasting with generative adversarial networks*. doi: 10
661 .20944/preprints202306.1492.v1
- 662 Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz Sabater, J.,
663 ... Thépaut, J.-N. (2020, 05). The era5 global reanalysis. *Quarterly Journal of*
664 *the Royal Meteorological Society*. doi: 10.1002/qj.3803
- 665 Imfeld, N., Pfister, L., Brugnara, Y., & Brönnimann, S. (2023, 03). A 258-year-long
666 data set of temperature and precipitation fields for switzerland since 1763. ,
667 19, 703–729. doi: 10.5194/cp-19-703-2023
- 668 Ivek, T., & Vlah, D. (2023, 01). Reconstruction of incomplete wildfire data using
669 deep generative models. *Extremes*, 26. doi: 10.1007/s10687-022-00459-1
- 670 Jia, X., Willard, J., Karpatne, A., Read, J. S., Zwart, J. A., Steinbach, M., & Ku-
671 mar, V. (2021, may). Physics-guided machine learning for scientific discovery:
672 An application in simulating lake temperature profiles. *ACM/IMS Trans.*
673 *Data Sci.*, 2(3). Retrieved from <https://doi.org/10.1145/3447814> doi:
674 10.1145/3447814
- 675 Kadow, C., Hall, D., & Ulbrich, U. (2020, 06). Artificial intelligence reconstructs
676 missing climate information. *Nature Geoscience*, 13. doi: 10.1038/s41561-020
677 -0582-5
- 678 Kashinath, K., Mustafa, M., Albert, A., Wu, J., Jiang, C., Esmailzadeh, S., ...
679 Prabhat, M. (2021, 04). Physics-informed machine learning: Case stud-
680 ies for weather and climate modelling. *Philosophical transactions. Series*
681 *A, Mathematical, physical, and engineering sciences*, 379, 20200093. doi:
682 10.1098/rsta.2020.0093
- 683 Kingma, D. P., & Ba, J. (2017). *Adam: A method for stochastic optimization*.
- 684 Lorenz, E. (1969, 07). Atmospheric predictability as revealed by naturally occurring
685 analogues. *J. Atmos. Sci.*, 26, 636–646. doi: 10.1175/1520-0469(1969)26<636:
686 APARBN>2.0.CO;2
- 687 Luterbacher, J., Xoplaki, E., Dietrich, D., Rickli, R., Jacobeit, J., Beck, C., ...
688 Wanner, H. (2002, 03). Reconstruction of sea level pressure fields over the
689 eastern north atlantic and europe back to 1500. *Climate Dynamics*, 18, 545–
690 561. doi: 10.1007/s00382-001-0196-6
- 691 NOAA National Geophysical Data Center. (2009). *ETOPO1 1 Arc-Minute Global*

- Relief Model*. NOAA National Centers for Environmental Information. (Accessed 03.08.2023)
- Pappert, D., Barriendos, M., Brugnara, Y., Imfeld, N., Jourdain, S., Przybylak, R., ... Brönnimann, S. (2022, 12). Statistical reconstruction of daily temperature and sea level pressure in europe for the severe winter 1788/89. *Climate of the Past*, 18, 2545-2565. doi: 10.5194/cp-18-2545-2022
- Pfister, L., Brönnimann, S., Schwander, M., Isotta, F., Horton, P., & Rohr, C. (2020, 04). Statistical reconstruction of daily precipitation and temperature fields in switzerland back to 1864. *Climate of the Past*, 16, 663-678. doi: 10.5194/cp-16-663-2020
- Pfister, L., Hupfer, F., Brugnara, Y., Munz, L., Villiger, L., Meyer, L., ... Brönnimann, S. (2019). Early instrumental meteorological measurements in switzerland. *Climate of the Past*, 15(4), 1345-1361. doi: 10.5194/cp-15-1345-2019
- Portenier, C. C., Lenggenhager, S., Schwander, M., Buck, A., & Foffa, S. (2017). The 1956 cold wave in western europe.
- Qian, W., Du, J., & Ai, Y. (2021, 01). A review: Anomaly-based versus full-field-based weather analysis and forecasting. *Bulletin of the American Meteorological Society*, 102, 1-52. doi: 10.1175/BAMS-D-19-0297.1
- Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., & Prabhat. (2019). Deep learning and process understanding for data-driven earth system science. *Nature*, 566(7743), 195-204. Retrieved from <https://doi.org/10.1038/s41586-019-0912-1> doi: 10.1038/s41586-019-0912-1
- Rössler, O., & Brönnimann, S. (2018). The effect of the tambora eruption on swiss flood generation in 1816/1817. *Science of The Total Environment*, 627, 1218-1227. doi: <https://doi.org/10.1016/j.scitotenv.2018.01.254>
- Schmutz, Y. (2024). *Werec3d v1.0.0*. Retrieved from <https://github.com/YannisSchmutz/WeRec3D/tree/v1.0.0> (Available at <https://github.com/YannisSchmutz/WeRec3D/tree/v1.0.0>) doi: 10.5281/zenodo.11262991
- Schultz, M., Betancourt, C., Gong, B., Kleinert, F., Langguth, M., Leufen, L., ... Stadtler, S. (2021, 02). Can deep learning beat numerical weather prediction? *Philosophical Transactions of The Royal Society A Mathematical Physical and Engineering Sciences*, 379. doi: 10.1098/rsta.2020.0097
- Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1), 60. Retrieved from <https://doi.org/10.1186/s40537-019-0197-0> doi: 10.1186/s40537-019-0197-0
- Stephenson, D., Wanner, H., Brönnimann, S., & Luterbacher, J. (2003, 01). The history of scientific research on the north atlantic oscillation. *Washington DC American Geophysical Union Geophysical Monograph Series*, 34, 37-50. doi: 10.1029/134GM02
- Sun, Q., Zhai, R., Zuo, F., Zhong, Y., & Zhang, Y. (2022, 02). A review of image inpainting automation based on deep learning. *Journal of Physics: Conference Series*, 2203, 012037. doi: 10.1088/1742-6596/2203/1/012037
- Taylor, K. (2001, 04). Summarizing multiple aspects of model performance in a single diagram. *Journal of Geophysical Research*, 106, 7183-7192. doi: 10.1029/2000JD900719
- Valler, V., Franke, J., Brugnara, Y., & Brönnimann, S. (2021, 05). An updated global atmospheric paleo-reanalysis covering the last 400 years. *Geoscience Data Journal*, 9. doi: 10.1002/gdj3.121
- Vaughan, A., Tebbutt, W., Hosking, S., & Turner, R. (2022, 01). Convolutional conditional neural processes for local climate downscaling. *Geoscientific Model Development*, 15, 251-268. doi: 10.5194/gmd-15-251-2022
- Wang, C., Haibin, H., Han, X., & Wang, J. (2019, 07). Video inpainting by jointly learning temporal structure and spatial details. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33, 5232-5239. doi:

- 747 10.1609/aaai.v33i01.33015232
748 Yao, Z., Zhang, T., Wu, L., Wang, X., & Huang, J. (2023, 03). Physics-informed
749 deep learning for reconstruction of spatial missing climate information in the
750 antarctic. *Atmosphere*, *14*, 658. doi: 10.3390/atmos14040658
751 Yu, F., & Koltun, V. (2016). *Multi-scale context aggregation by dilated convolutions*.