Norwegian Meteorological Institute

METreport

No. 04/2020
ISSN 2387-4201
Free

ClimNorm

- temperature data set, gap filling methods and regional analysis to prepare new climate normals

Ole Einar Tveito[1], Svetlana Aņiskeviča[4], John Cappelen[6], Erik Engström[2], Herdis Motrøen Gjelten[1], Caroline Drost Jensen[6], Pauli Jokinen[3], Elinah Khasandi Kuya[1], Cristian Lussana[1], Antti Mäkelä[3], Kaupo Mändla[5], Kairi Vint[5], Lennart Wern[2], Viesturs Zandersons[4]

[1]Norwegian Meteorological Institute, [2]Swedish Meteorological and Hydrological Institute, Sweden, [3]Finnish Meteorological Institute, Finland, [4]Latvian Environment, Geology and Meteorology Centre (LVGMC), Latvia, [5]Estonian Environment Agency, [6]Danish Meteorological Institute, Denmark

Norwegian Meteorological Institute

METreport

Title	Date
ClimNorm - temperature data set, gap filling methods and regional analysis to prepare new climate normals	June 29, 2020
Section Division for Climate Services	**Report no.** 04/2020
Author(s) Ole Einar Tveito, Svetlana Aniskeviča, John Cappelen, Erik Engström, Herdis Motrøen Gjelten, Caroline Drost Jensen, Pauli Jokinen, Elinah Khasandi Kuya, Cristian Lussana, Antti Mäkelä, Kaupo Mändla, Kairi Vint, Lennart Wern, Viesturs Zandersons	**Classification** ● Free ○ Restricted
Client(s) NORDMET/NFCS	**Client's reference**

Abstract

ClimNorm is a network activity under the framework of NORDMET/NFCS that aims to support the NMHSs in the Nordic Region in their efforts to calculate new climate normals for the 1991-2020 period by exchanging ideas, experiences and algorithms for data gap filling, homogenisation and dissemination. Central in this activity is to collect and share monthly precipitation and temperature data from the region. This report presents the joint temperature data sets, an overview of possible gap filling methods and some effects of homogenisation of temperature data.

Keywords

Climate normals, temperature, gap filling, homogenisation

Disciplinary signature
Hans Olav Hygen

Responsible signature
Cecilie Stenersen

Contents

1 Introduction

Climate normals are defined by WMO (*WMO*, 2017) as 30-year representative averages of climate variables referring to the most recent 30-year period finishing in a year ending with 0. This definition replaces the previous definition of consecutive non-overlapping 30 year periods (1901-30, 1931-60, 1961-90 and the upcoming 1991-2020).

Calculation of the climate normals is, given that the input data series are complete and of good quality, a straightforward procedure. But the reality is that many series are incomplete and/or inhomogeneous. Since the normals are sensitive to the averaging period, efforts have to be made to secure a robust and consistent basis for calculations of climate normals.

In all Nordic countries the observation network has been drastically changed over the last 15-30 years. This has caused challenges for the calculation of climate reference values (the normals) as they require complete and preferably homogeneous data series. To be able to calculate representative climate normals efforts has to be taken to (i) fill in gaps in incomplete data series and (ii) assess and adjust for inhomogeneous and inconsistent data series.

The Nordic region is characterized by large variations and gradients in weather and climate caused by different topographical and coastal influences. The areas along the national borders are sparsely covered by meteorological stations, and data exchange across these borders will provide a better data basis for calculation of stations normals in the region. The climate services in the Nordic countries have a long and profound history of collaboration within climatology. Organized under the Nordic Framework for Climate Services (NFCS), ClimNorm will be a continuation of this tradition, focusing on establishing a high quality homogenized reference data set, evaluating gap filling methods and assessing spatial and temporal trends and variability, producing a Nordic climate atlas for the 1991-2020 normal period.

This report describes the compilation of a pan-Nordic temperature observation data set that will form the basis for activity. It describes the data set, and the challenges with incomplete data. Further are some possible methods for filling gaps in time series pre-

sented and discussed. A few examples of applying and comparing some of these methods are also presented. Finally are some effect of homogenisation highlighted.

2 Data set

2.1 Observation data

The first ClimNorm data set contains long term monthly temperature data series from the Nordic region during the period 1901 until near present (2018). It is compiled applying data from the national meteorological services in Latvia, Estonia, Finland, Sweden, Denmark and Norway. Figure 1 shows the location of the available temperature observation series. Figure 2 shows the number of stations with observations each year during the period 1901-2018. In the early part of the last century the number of available observations is rather low. Since around 1960 the number of observations has increased with a maximum around 1970. Then there was a decrease until 2000. During the last 20 years the number of observations has increased considerably, especially due to automation and access to observations from organisations and governmental institutions other than the national meteorological services. The number of series with complete or almost complete records for the entire 1961-2018 period are however relatively few compared to the total number stations that have been in operation during this period (Figure 3).

Most of the analyses in ClimNorm will demand complete data series. Figure 4 shows the distribution of series with 30 consecutive years of observations within the period 1901-2018, showing that there are not too many series that covers the entire period. It is however quite obvious that many of the shorter series are due to relocations, and that merging of data series for such stations should make it possible to increase the number of series with complete, or almost complete, data coverage. For series with gaps, gap filling methods could be applied to fill in the missing data. In the next sections, the steps to complete observation series are described and discussed, they include the merging of short series and the application of different approaches for filling gaps in data series.

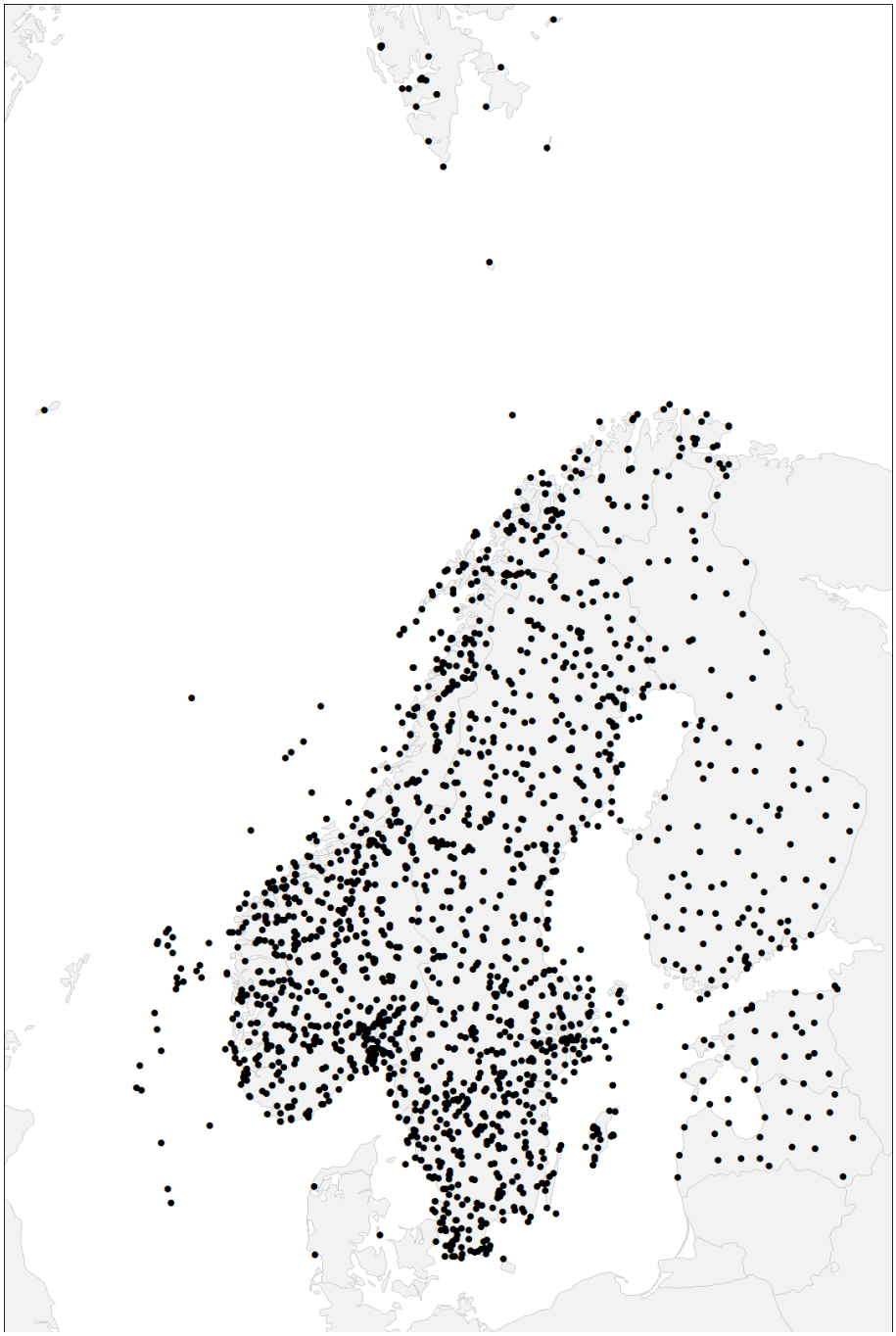


Figure 1: Locations of all the stations in the ClimNorm temperature data set

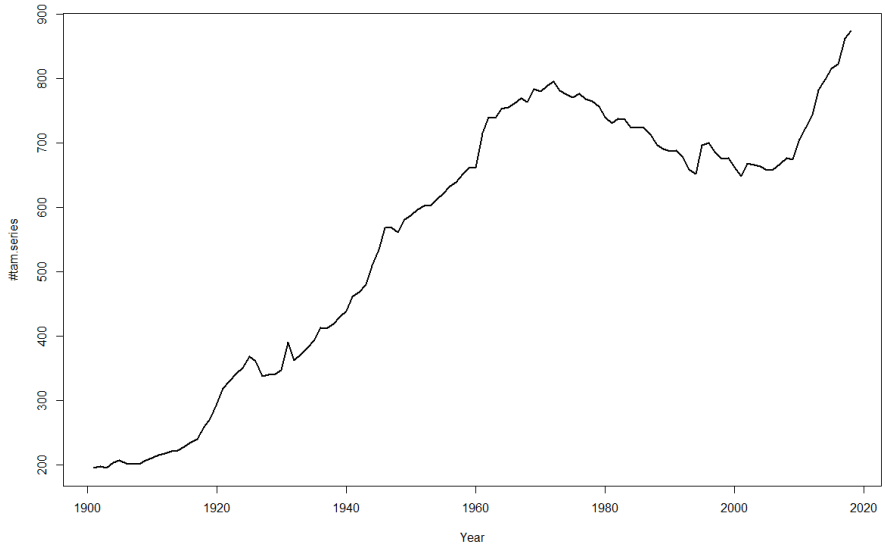


Figure 2: Number of stations pr. year

2.2 Merging data series

When considering stations that have been relocated there is a possibility to generate longer observation time series. We have merged series according to the following principles:

- horizontal distance <10 km
- maximum vertical distance $\pm$ 100 m
- partial series should have a substantial length. How long must be considered in each case. (subjective assessment)

The merging was carried out semi-automatically. First, all series having more than 10 observation years within the 1961-2018 period were automatically scanned to identify neighbouring series fulfilling the requirements listed above. The target series were plotted together with the potential partner series. Figure 6 shows the connection between the current operational weather station 47024890 Nesbyen, Norway and the historical series from the same neighbourhood.

These plots were manually inspected and used to match series that could be merged. No series could be used in more than one final series. The series having the most recent observations are kept as the target for the final series, and the merged series keep

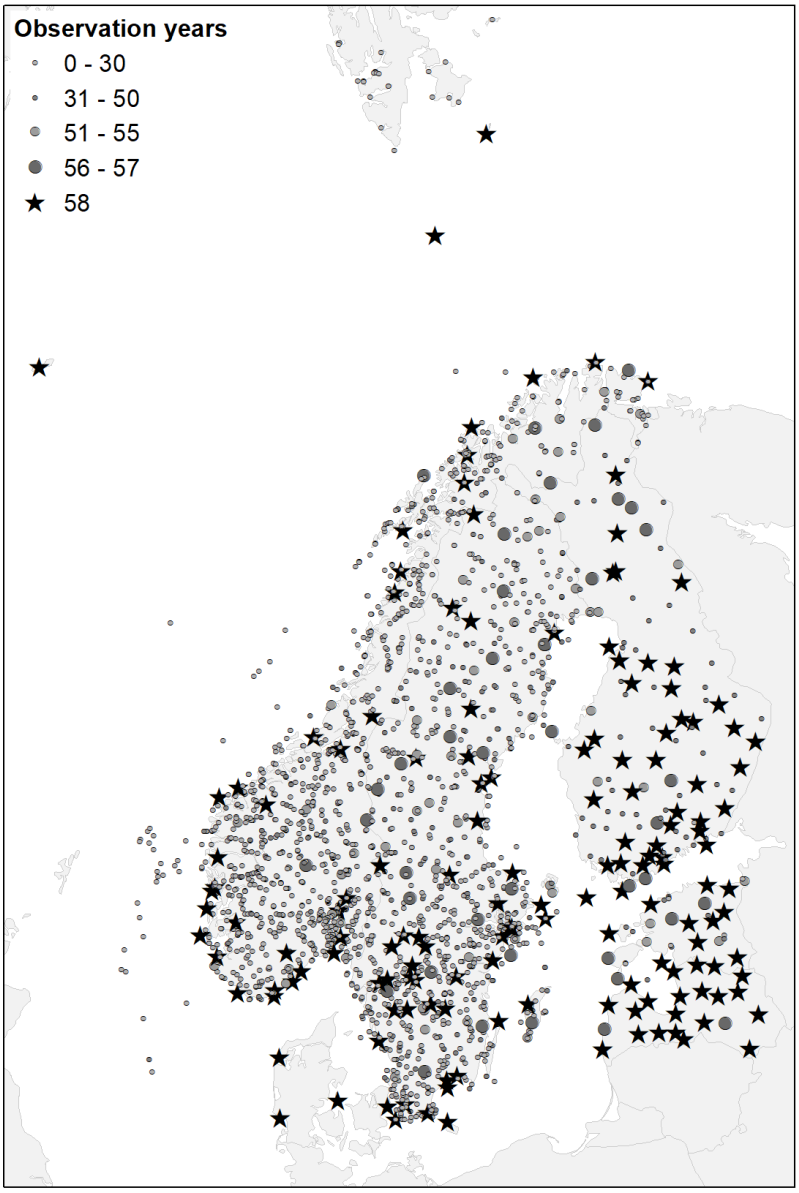


Figure 3: Data coverage 1961-2018.

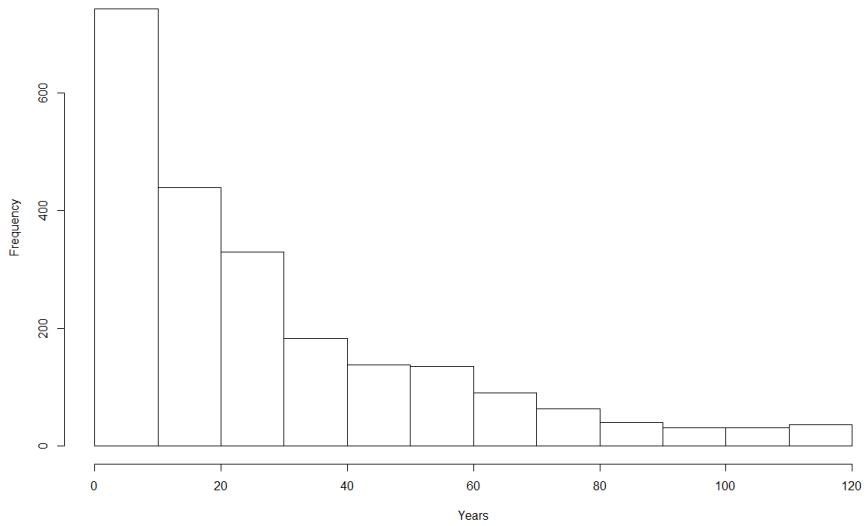


Figure 4: Frequency of observation series lengths for the period 1901-2018

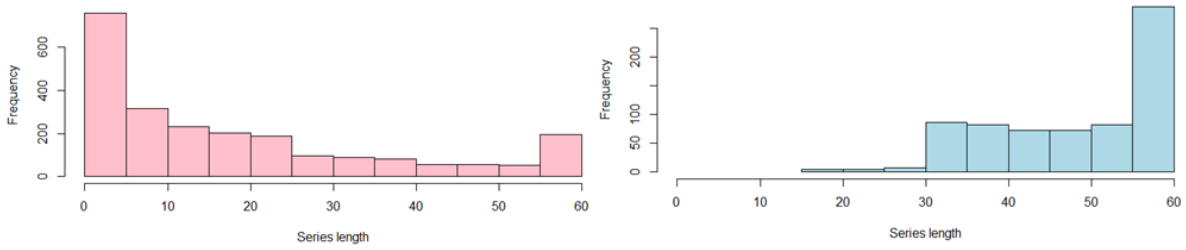


Figure 5: Frequency of observation series lengths in the period 1961-2018, before (left) and after (right) merging of station series

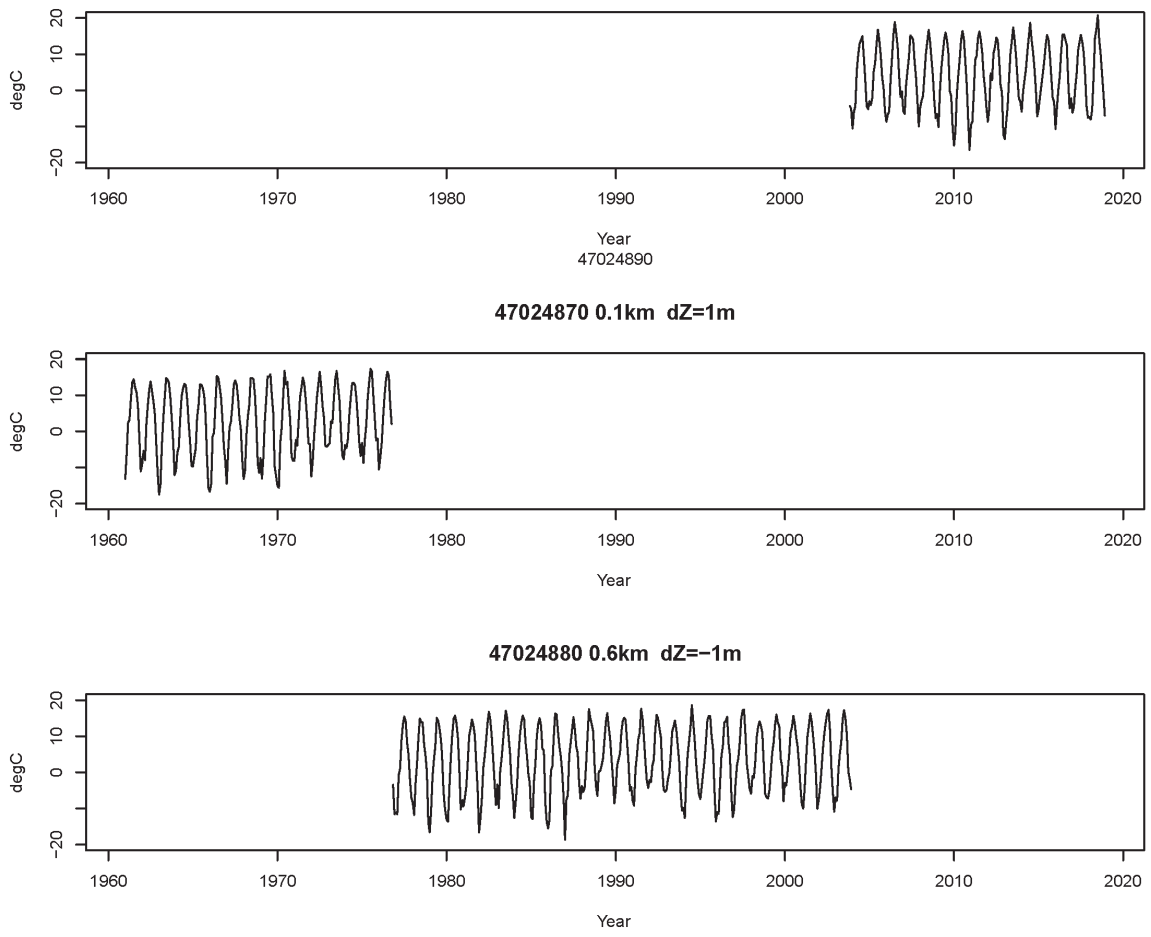


Figure 6: Visualisation of the series used to merge the series 47024890 Nesbyen, Norway

the Name/ID of the target series. After the merging procedure 693 series remained for the further analyses and method testing. Figure 5 shows the distribution of series length before and after merging in the period 1961-2018.

3 Filling gaps in time series

Climatic time series are often incomplete due to a number of reasons: break in communication, technical errors at the station, temporary close down, relocation of observation locations, etc. It is therefore often a need to fill missing values in these time series to establish a robust, consistent and possibly homogeneous data basis for addressing climate variability and change. In this report, we present different methods that can be used to fill gaps in time series. The short review presented here is based on a literature survey as well as experiences gained applying different methods within the Nordic countries.

Filling gaps is in principle data estimation, and often an analysis of the existing time series is needed. Many of the general principles and main methods applied to analyse climate data are presented in *WMO* (2018). We have to differentiate between *interpolation* and *extrapolation* of data. When data exists both before and after the gap we *interpolate* data. If gaps are filled in one of the ends of the time series only, data are *extrapolated*.

There are a number of different approaches that can be applied for filling gaps in time series, and they can coarsely be grouped into the following groups:

- Spatial neighbourhood
- Spatial interpolation
- Statistical methods
- Downscaling methods

In the following, these different approaches will be presented and discussed. Some examples on how they might perform on monthly time series in Fennoscandia will also be presented and discussed.

3.1 Spatial interpolation methods

This group of methods assumes that geographical vicinity can be applied for filling gaps in time series. The methods that fall in under this category are

- Nearest neighbour (Thiessen/Voronoï)
- Triangulation
- Inverse distance weighting
- Geostatistical methods

The common denominator of these methods is that they take the spatial distance into consideration when an interpolated value is assigned.

3.1.1 Nearest neighbour

This is the simplest method. The interpolated value is taken from the nearest observation with a valid value. Thiessen polygons or the Voronoï diagram method are algorithms that can be applied to find the nearest data point. These two methods are theoretically similar, but are developed and applied within different communities. The Thiessen polygon method has been widely applied to estimate areal precipitation for engineering and hydrological applications, while the Voronoï approach is more common within mathematics and computer science.

In these algorithms every point in space is assigned to the nearest observation point, forming a polygon (area) around each observation that will be assigned with the same value as the observation point. The result will become a discontinuous surface. See an example in figure 7.

3.1.2 Averaged neighbour anomalies

A rather simple method based on averaging neighbouring values, normalized by dividing them with their respective mean values. This method is applied among others by the *climatol* package (*Guijarro*, 2018).

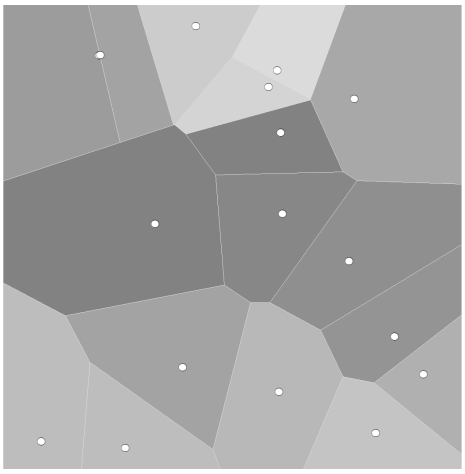


Figure 7: Nearest neighbour estimation (Thiessen/Voronoï)

3.1.3 Triangulation

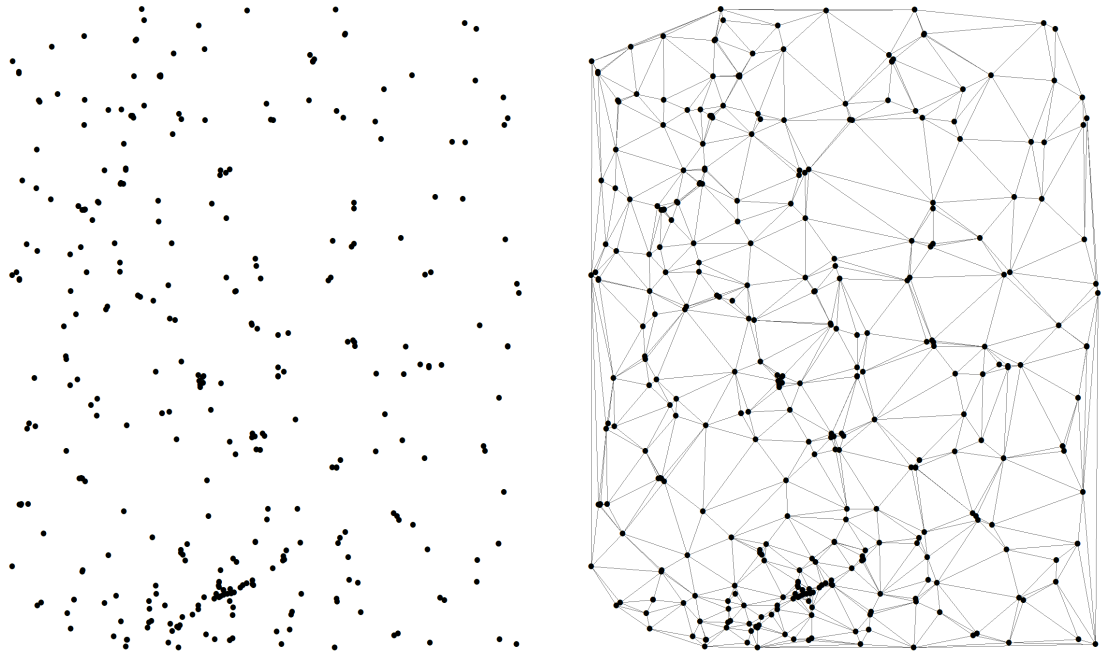
Triangulation is strongly related to the Voronoï diagram method. Triangulation is in fact used to determine the polygons. But instead of making discrete polygons, triangulation creates a continuous faceted surface built up by triangles between the observation points (Figure 8).

The slope of the surface along with the distance to the three corners will give an interpolation value at any point within the triangle. The estimate $\hat{z}_D$ in point D shown in figure 9 will then be a combination of the slope of the surface and the values at the corners A, B and C as expressed in equation 1.

$$\begin{aligned} \alpha &= (x_C - x_A)\,(y_B - y_A) - (x_B - x_A)(y_C - y_A) \\ z_{dx} &= ((y_B - y_A)(z_C - z_A) - (y_C - y_A)(z_B - z_A))/\alpha \\ z_{dy} &= ((x_C - x_A)(z_B - z_A) - (x_B - x_A)(z_C - z_A))/\alpha \\ \hat{z}_D &= z_A + (x_D - x_A)z_{dx} + (y_D - yA)z_{dy} \end{aligned} \tag{1}$$

3.1.4 Inverse distance weighting

Inverse distance weighting (IDW) is often included in the group of statistical spatial interpolation methods despite the fact that it only takes the distance to the n nearest observation points into consideration. The estimate is a linear combination of the n nearest observations where the weight of each observation is proportional to the inverse of the distance



(a) Observation points (b) TIN

Figure 8: Establishing a triangular network (TIN) from point locations.

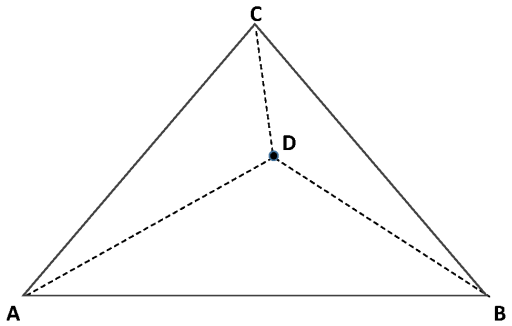


Figure 9: Estimation by triangulation

between the prediction point and the observation points.

$$z_0 = \frac{\sum_{i=1}^{n} w_i(x) u_i}{\sum_{i=1}^{n} w_i(x)} \tag{2}$$

where

$$w_i x(i) = \frac{1}{d(x_0, x_i)^p} \tag{3}$$

x_0 is the prediction point, while u_i are the observation points. d denotes the distance between x_0 and x_i, and p is the power of the distance weighting. Higher p-values give higher weight to the closer points, lower value will give a larger influence of the more distant observation points. The most commonly used p-value is 2.

When applying IDW it is possible to adjust for anisotropies (trends) in data, as well as defining a proper neighbourhood. A disadvantage of IDW, which is also the case for other spatial neighbourhood methods, is that it is limited to provide estimates within the observed minimum and maximum value interval at the observation stations.

3.1.5 Geostatistical spatial interpolation

The geostatistical approach differs from the methods described above since they also consider characteristics of the observed field in addition to the geometric features of the observation points. The main assumption for this type is that they assume spatial covariance structure, where the covariance depends on the distance (h) between the observations. To this group of methods we can count in kriging and optimal interpolation (OI). OI is a well known concept in atmospheric sciences as the fundamental principle for data assimilation. The problem was formulated by *Eliassen* (1954) and further developed by *Gandin*

and Hardin (1965), and therefore often referred to as Gandin interpolation. Kriging is a similar approach developed for mining applications. The spatial covariance structure applied in OI is a correlogram, while kriging applies a semivariogram. Kriging is developed for processes where the number of realisations are few, or even only one, while OI takes advantage of multiple realisations. *Creutin and Obled* (1982) gives a nice overview of several spatial interpolation methods, showing that even if they are based on different theories and assumptions they can be broken down to similar mathematical expressions.

3.2 Statistical methods

Statistical methods are often used to model timeseries, and can thus be applied also for filling gaps in time series. Statistical models will only consider the statistical properties, and will not consider the geographical locations (unless these are included in the predictor fields as described by e.g. *Tveito* (1998)).

3.2.1 Linear regression methods

Linear regression is maybe one of the most applied approaches to model a process, and to fill time series. It is basically a line fitting technique, assuming an underlying gaussian distribution. Mathematically it can be expressed as

$$\hat{X} = a + bY \tag{4}$$

When more predictors (e.g. other times series) are applied the mathematical expression takes this form:

$$\hat{X} = a + \sum_{i=1}^{n} b_i Y_i \tag{5}$$

When filling gaps in time series $\hat{X}$ the predictors Y_i are usually other time series. The selection of predictor series is normally based on one, or a combination of these criteria:

- Covariance
- Distance
- Representative neighbours

3.2.2 Principal component analysis

One possible concept to fill gaps in time series is to apply principal component analysis (PCA). This method, also often referred to as empirical orthogonal functions (EOF), is in principle a data reduction method used to identify spatial and temporal patterns in a data set (e.g. a set of time series). Even though it is regarded as a multivariate statistical method, it does not assume an underlying statistical distribution nor does it depend on any statistical tests. The components (or functions) are found by calculating the eigenvectors and eigenvalues of the anomaly covariance matrix. That means that the input data are normalized before the eigenvalues/eigenvectors are calculated. The eigenvalues show the explained variance of each mode (eigenvectors) of the input data. The principle applying PCA for gap filling is a bit similar to the linear regression approach, but the predictors are PCs instead of time series. PCA is a multivariate statistical method where a set of correlated data series are transformed into a set of orthogonal (and thus uncorrelated) principal components that describe the total variance in the original time series. Mathematically this can be expressed as:

$$X'_k(t) = \sum_{i=1}^{m} h_{ik}\beta_i(t) \tag{6}$$

where $X'_k(t)$ is an observation series, $\beta_i(t)$ are the temporal principal components describing variations in time. h_{ik} are the spatial components that weight the $\beta_i(t)$ according to their influence in the series. The weight factors h will thus give a spatial signal of the influence of the individual components. i denotes the component number, and k is the location index. The $\beta_i(t)$'s are sorted meaning that the one explaining most of the total variance is the first, the second most second etc. Since most of the original total variance is explained by the first principal components, this method can be applied for reducing the dimensionality in large correlated dataset. Thus a few PCs will be applied to explain most of the variance in the target series and be used to estimate missing data. The sum of all weighted components will reproduce the original time series.

The principal components $\beta_i(t)$ describes the temporal variations. For each time series these components are weighted h_{ik} .

Hisdal and Tveito (1993) showed that these weight coefficients can be estimated as:

$$h'_{ik} = \frac{\rho(x'_k(t')b_i(t'))}{\sigma(\beta_k(t'))} \tag{7}$$

This relation can be applied to fill gaps in incomplete time series by calculating h'_{ik} from the common period between the time series and the principal components. The final series is derived from

$$\hat{X}_k = \sum_{i=1}^{n} h'_{ik} \beta_i(t) \tag{8}$$

The PCA concept can thus be applied to reconstruct dataseries in data sparse periods based on periods having good data coverage given that a joint calibration period between the high resolution and lower resolution datasets exist (*Schiemann et al.*, 2010). *Hisdal and Tveito* (1993) also showed that regionalization of the data series into regions (or classes) with the same characteristics gave more robust principal components with regard to gap filling, since all components then will contain significant information for the series to be completed. The h_{ik}'s often show a spatial pattern, and are thus representations of regional climate signals. *Hisdal and Tveito* (1992) demonstrated that the h_{ik}'s can be estimated applying a spatial interpolation method (e.g. kriging) to estimate timeseries at locations without calibration data.

3.2.3 Testing PCA as gapfilling method

By nature, the EOFs are not correlated, and can therefore be treated as independent predictors. The EOFs are quite often calculated from timeseries with the same climatological characteristics as the candidate series. Therefore a classification of the temperature series into 12 regions where carried out by a k-means cluster analysis. The twelve classes formed geographical regions as shown in Figure 10.

The PCA was carried out both as a global analysis, including all 693 series and as an analysis of each of the twelve regional data sets. In order to avoid impact of the annual cycle signal in the temperature series, the input series are normalised by the monthly mean values. The PCAs are also performed for each of the months individually. This means that the twelve PC-analyses are carried out for each region. Tables 1, 2 and figure 11 shows that the regional PCs explain more of the variance by the first few components than the global analysis for most regions. The exception is region 5 (dark green line) which includes the Arctic series. For this region more components (4-5) are needed to cover 95 % of the variance than for the other regions.

Table 1: Proportion of variance explained by the first principal component

Region	Jan	Feb	Mar	Apr	May	June	Jul	Aug	Sep	Oct	Nov	Dec
Global	0.800	0.841	0.833	0.708	0.711	0.674	0.732	0.738	0.776	0.729	0.732	0.797
Region 1	0.957	0.952	0.958	0.944	0.934	0.937	0.93	0.921	0.945	0.947	0.941	0.950
Region 2	0.926	0.936	0.959	0.951	0.949	0.911	0.905	0.914	0.958	0.948	0.923	0.947
Region 3	0.935	0.929	0.908	0.882	0.826	0.833	0.877	0.902	0.928	0.897	0.925	0.945
Region 4	0.974	0.980	0.972	0.943	0.960	0.960	0.953	0.938	0.971	0.972	0.972	0.978
Region 5	0.766	0.679	0.790	0.785	0.758	0.688	0.688	0.691	0.801	0.793	0.748	0.743
Region 6	0.981	0.981	0.974	0.916	0.922	0.926	0.957	0.933	0.961	0.959	0.960	0.975
Region 7	0.958	0.964	0.973	0.962	0.957	0.957	0.957	0.953	0.973	0.976	0.956	0.973
Region 8	0.939	0.953	0.950	0.919	0.922	0.912	0.938	0.940	0.926	0.922	0.909	0.941
Region 9	0.955	0.961	0.966	0.937	0.927	0.934	0.939	0.944	0.959	0.967	0.948	0.956
Region 10	0.953	0.966	0.961	0.936	0.919	0.924	0.926	0.932	0.936	0.938	0.936	0.949
Region 11	0.967	0.966	0.960	0.925	0.881	0.855	0.927	0.936	0.920	0.931	0.926	0.956
Region 12	0.981	0.978	0.965	0.919	0.908	0.931	0.954	0.939	0.961	0.968	0.976	0.980

Table 2: Proportion of variance explained by the first five principal components

Region	Jan	Feb	Mar	Apr	May	June	Jul	Aug	Sep	Oct	Nov	Dec
Global	0.961	0.963	0.962	0.934	0.926	0.921	0.943	0.939	0.949	0.950	0.936	0.960
Region 1	0.991	0.991	0.991	0.987	0.991	0.992	0.992	0.990	0.990	0.990	0.985	0.990
Region 2	0.991	0.992	0.994	0.990	0.993	0.993	0.994	0.988	0.993	0.993	0.990	0.993
Region 3	0.985	0.985	0.981	0.975	0.968	0.970	0.975	0.981	0.983	0.976	0.982	0.987
Region 4	0.997	0.997	0.995	0.991	0.992	0.992	0.992	0.985	0.991	0.996	0.997	0.997
Region 5	0.982	0.981	0.982	0.979	0.968	0.968	0.964	0.958	0.983	0.985	0.982	0.985
Region 6	0.997	0.997	0.993	0.983	0.983	0.984	0.989	0.982	0.988	0.990	0.994	0.996
Region 7	0.998	0.998	0.998	0.996	0.997	0.996	0.998	0.994	0.996	0.998	0.998	0.998
Region 8	0.989	0.991	0.988	0.979	0.980	0.978	0.981	0.982	0.978	0.984	0.984	0.988
Region 9	0.991	0.993	0.992	0.985	0.983	0.987	0.987	0.985	0.989	0.992	0.991	0.992
Region 10	0.990	0.993	0.992	0.986	0.985	0.984	0.986	0.983	0.984	0.985	0.988	0.988
Region 11	0.993	0.993	0.988	0.977	0.968	0.958	0.977	0.974	0.973	0.978	0.983	0.990
Region 12	0.998	0.998	0.995	0.989	0.987	0.988	0.990	0.987	0.992	0.994	0.996	0.997

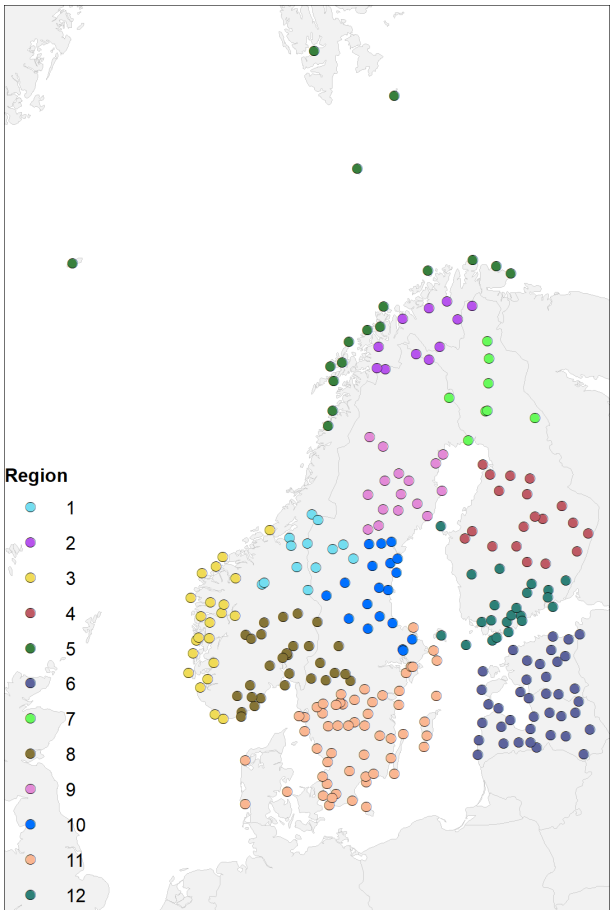


Figure 10: Twelve temperature series classes defined by a k-means cluster analysis

The analyses are validated by a cross-validation. Each series is extracted from the input data set before the PCA is carried out for the remaining series. The weight coefficients are estimated using 1961-90 as calibration period and 1991-2018 as verification period. Figure 12a shows the root mean square error (RMSE) for the global analysis. The average global RMSE is 0.37, minimum is 0.21 and maximum is 1.15. For the regional analysis the mean is 0.23, the minimum 0.11 and the maximum 0.91. This clearly indicates that the regional analysis provides more exact estimates than the global analysis. Figure 13 shows that except for one station (358101464) located in the Gulf of Bothnia, all stations achieve lower RMSE when regional PCAs are applied compared to the global. Figure 14 shows the typical behaviour at a single station. Values predicted by global PCA are more scattered than the ones estimated by the regional PCs. They are closer to the straight line indicating a perfect match. The maps in figure 12 confirm this. These maps also show that the lowest RMSEs, and thus the best results, are obtained in the south eastern parts of the study domain, in southern Sweden, Finland and the Baltic countries. These are areas

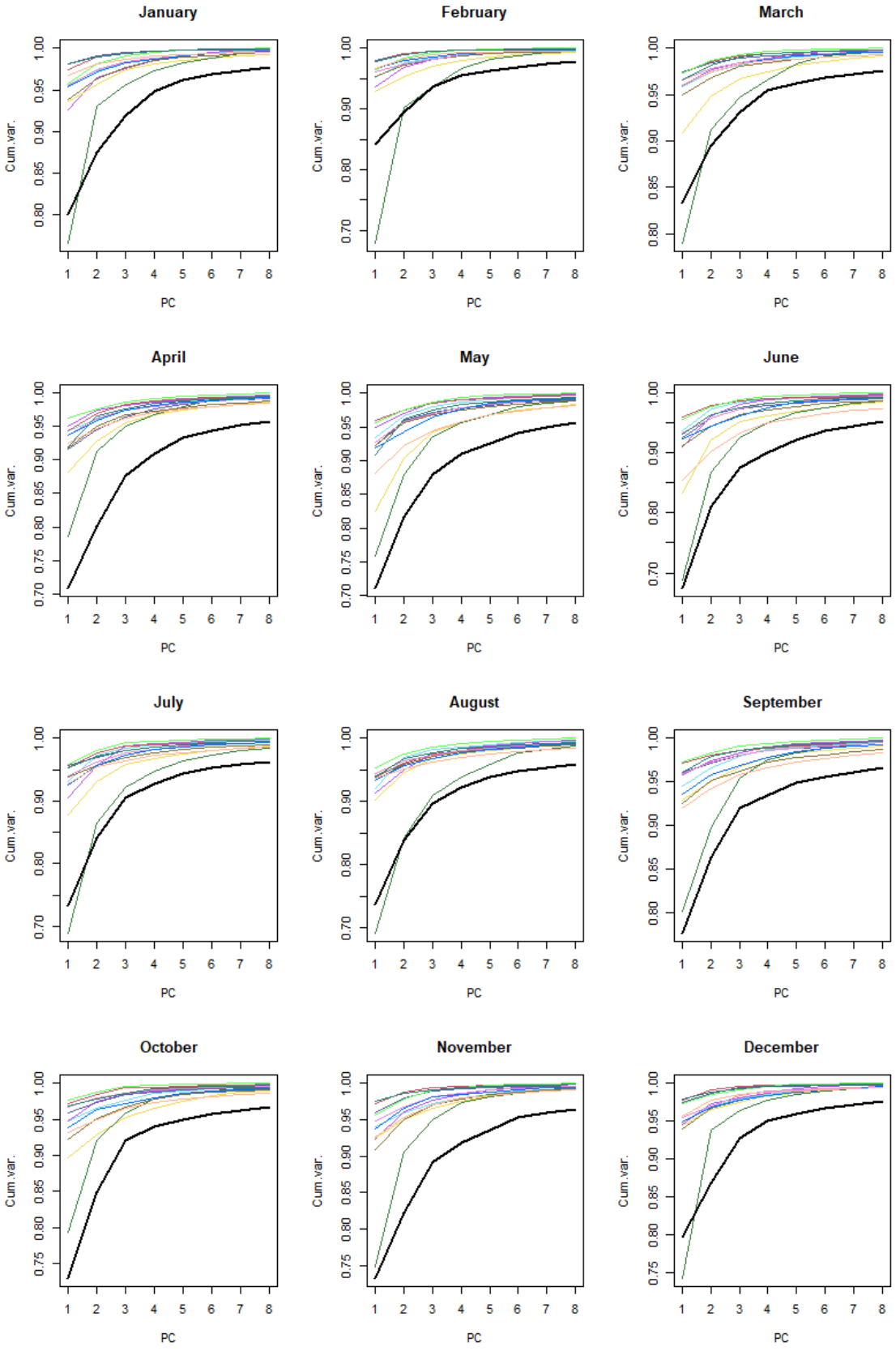
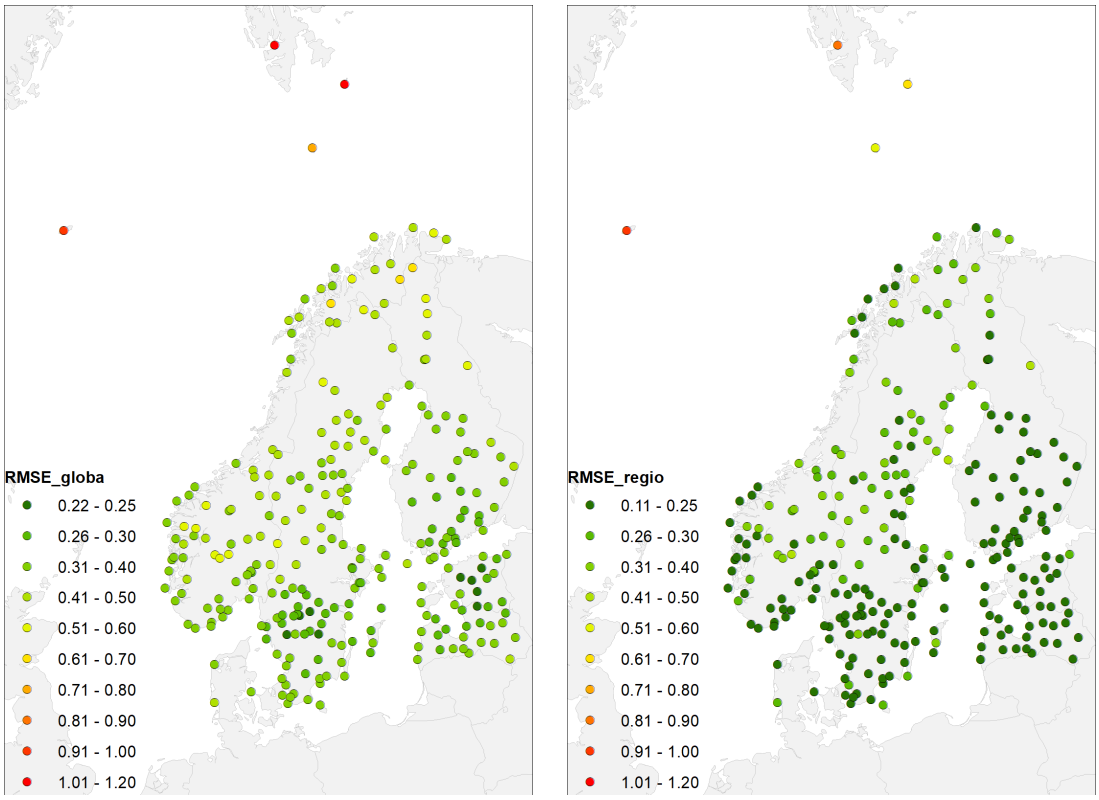


Figure 11: Cumulative proportion of variance explained by the first eight PCA's for the global (thick black line) and the twelve regional PCAs (coloured, see fig. 10) for each month

with quite high station density and relatively small topographical variations.

The largest RMSE values are found for the mountain regions and northern parts of the study area. The absolutely highest RMSEs are for the Arctic stations, where the distances between the stations are huge.



(a) Global PCA (b) Regional PCA

Figure 12: RMSE of estimation of timeseries applying PCA.

Figure 12b shows the performance for the regional analysis. The higher number of dark green dots across the study domain show that the RMSE is lower, and that the regional analysis improves the estimates especially in the western and northern parts of the domain where station density is low. The climatic characteristics in these areas are suppressed in the global analysis due to the more numerous observation series in the south-eastern part of the study domain. For the Arctic series the RMSE are reduced, but still quite large.

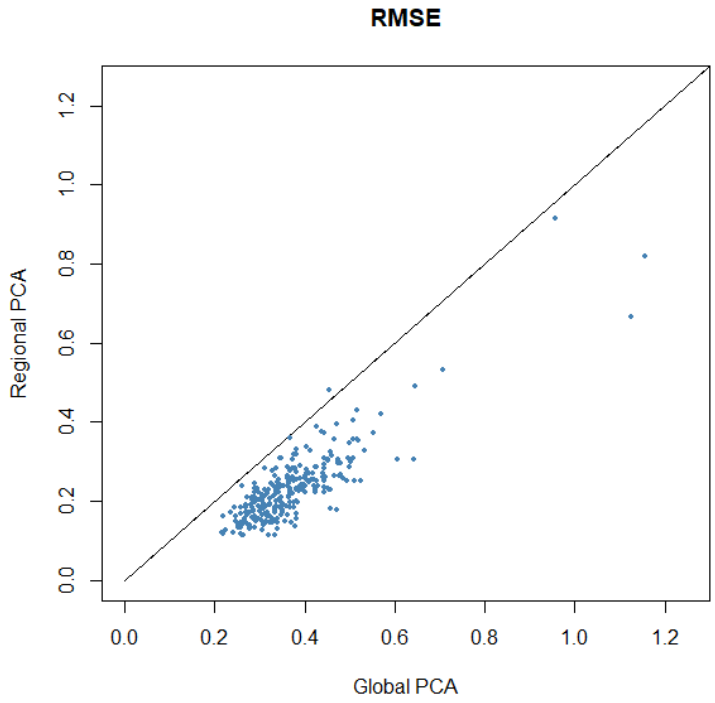


Figure 13: RMSE of global PCA vs regional PCA

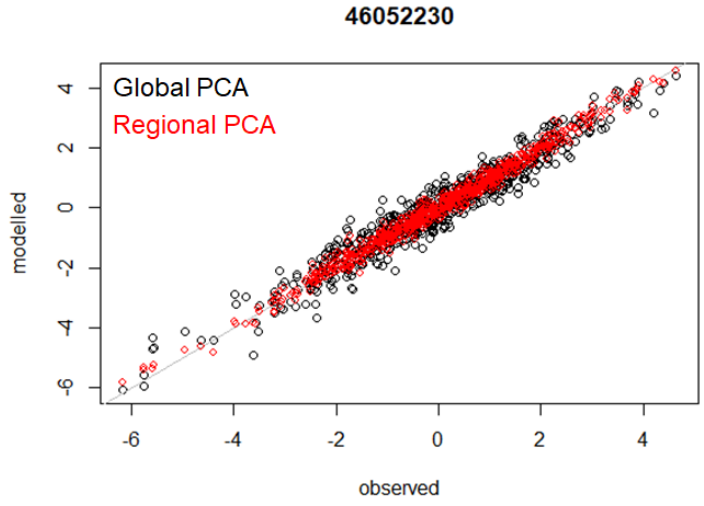


Figure 14: Observed vs predicted values by global and regional PCA at station 46052230

3.3 Gap filling applying gridded data

Grid-based interpolation is a method to calculate values in places where one has no observations. At MET as well as many other data providers, spatial interpolation methods have been developed which are used to produce observation-based gridded data (*Tveito et al.*, 2000, 2005; *Engeset et al.*, 2004; *Lussana et al.*, 2018a,b). Such algorithms can also be used to calculate point values, as described above. The advantage of such observation based gridding is that they often provide data with a very high spatial resolution and that they cover quite long time periods (several decades).

Gridded data sets offer a unique possibility to achieve complete data records. These data sets are complete in terms of spatial and temporal coverage. There are several approaches to establish gridded data. The most used in climatology are either observation gridding where the major data sources are in-situ observations of the climate element of interest. In the Nordic countries FMI (*Aalto et al.*, 2016), DMI (*Wang and Scharling*, 2010) and MET Norway all produce such data. One pan Nordic dataset of this type currently exists, the Nordic Gridded Climate Dataset NGCD (*Lussana et al.*, 2018a; *Tveito et al.*, 2005) that is a part of the C3S surface climate observation monitoring service. NGCD is produced by MET Norway and is updated every 6 months. The advantage of observation gridding is, given that the number of input data points is sufficient and relatively evenly distributed, that they provide quite precise data at a very high spatial resolution. As an example, NGCD provides daily values of temperature and precipitation at a spatial resolution of 1 x 1 km. The main disadvantages are that the physical consistency between parameters are lost, and that data are extrapolated outside the input data sample domain.

The other dominating approach is the use of atmospheric models for reanalysis or hindcasts. These models provide physically consistent data but often at lower spatial resolution due to the heavy computer demands. State of the art global reanalysis such as (*Hersbach et al.*, 2020; *C3S*, 2017) has a spatial resolution around 30 km. Regional reanalyses like CERRA (Europe) and CARRA (Arctic) that are under development will provide data with 6 km or less resolution. Some reanalyses provide ensemble data, mostly with a lower resolution than the deterministic product.

When applying gridded data for gap filling the bias between observations and the estimates from the gridded data set needs to be corrected. A common way to do this is to apply a linear regression model. This is maybe one of the most applied approaches to model a process, and to fill time series. It is basically a line fitting technique, assuming an underlying gaussian distribution. Mathematically it can be expressed as

$$X = a + bY \tag{9}$$

where X is the observation, Y is the predictor (in this case the grid model value) and a and b are model parameters. This model describes the relation as a straight line. It can be further developed to describe a curve by

$$X = a + bY + cY^2 \tag{10}$$

To demonstrate this we have fitted data for the 621 ClimNorm stations that are inside the NGCD coverage by downscaling ERA5 and NGCD type values. Figure 15 shows the histograms of RMSE of the downscaled mean daily temperatures at these locations. It is easy to see that (i) the raw NGCD estimates are more precise than the ERA5 values. This should be expected since the input data to NGCD to a large extent are the same as in the ClimNorm dataset. The spatial distribution of the RMSE of the raw estimates 16a and 16d shows that both methods provide quite precise estimates in areas with small topographical variations. Along the Norwegian west coast and in the mountains the RMSE is considerable, especially for ERA5. The NGCD2 estimates have lower RMSE, mainly because the gridded data set has a high spatial resolution (1x1 km) and that it is based on observations only. Both data sets show a bias, and the bias adjustment, both linear (16b and e) and polynomial (16c and f), improves the estimates. For ERA5 is there a large improvement. Applying a polynomial fit gives estimates that are slightly more precise than the linear fit. For NGCD2 fitting a linear model also leads to a real improvement of the estimates, and the polynomial fit shows slightly better estimates than the linear fit. One possible danger applying a polynomial function to adjust biases that values at the end of the distribution function is that the fitted curve might deviate from the true distribution due to undersampling at the end of the tails of the distribution compared with the central part of the distribution (extrapolation effect).

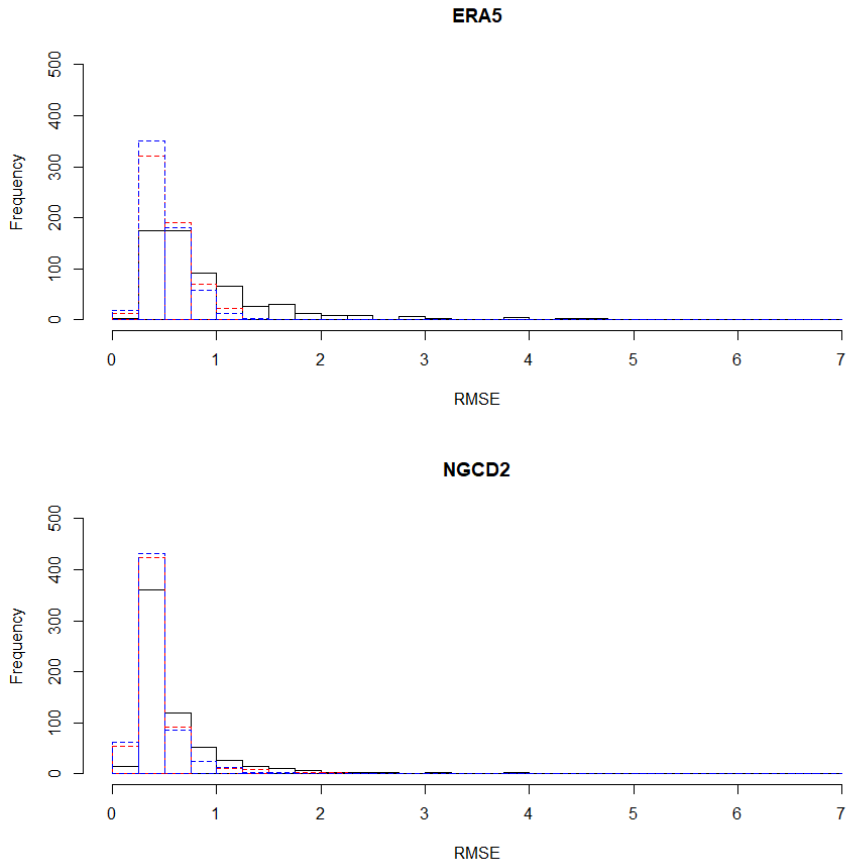
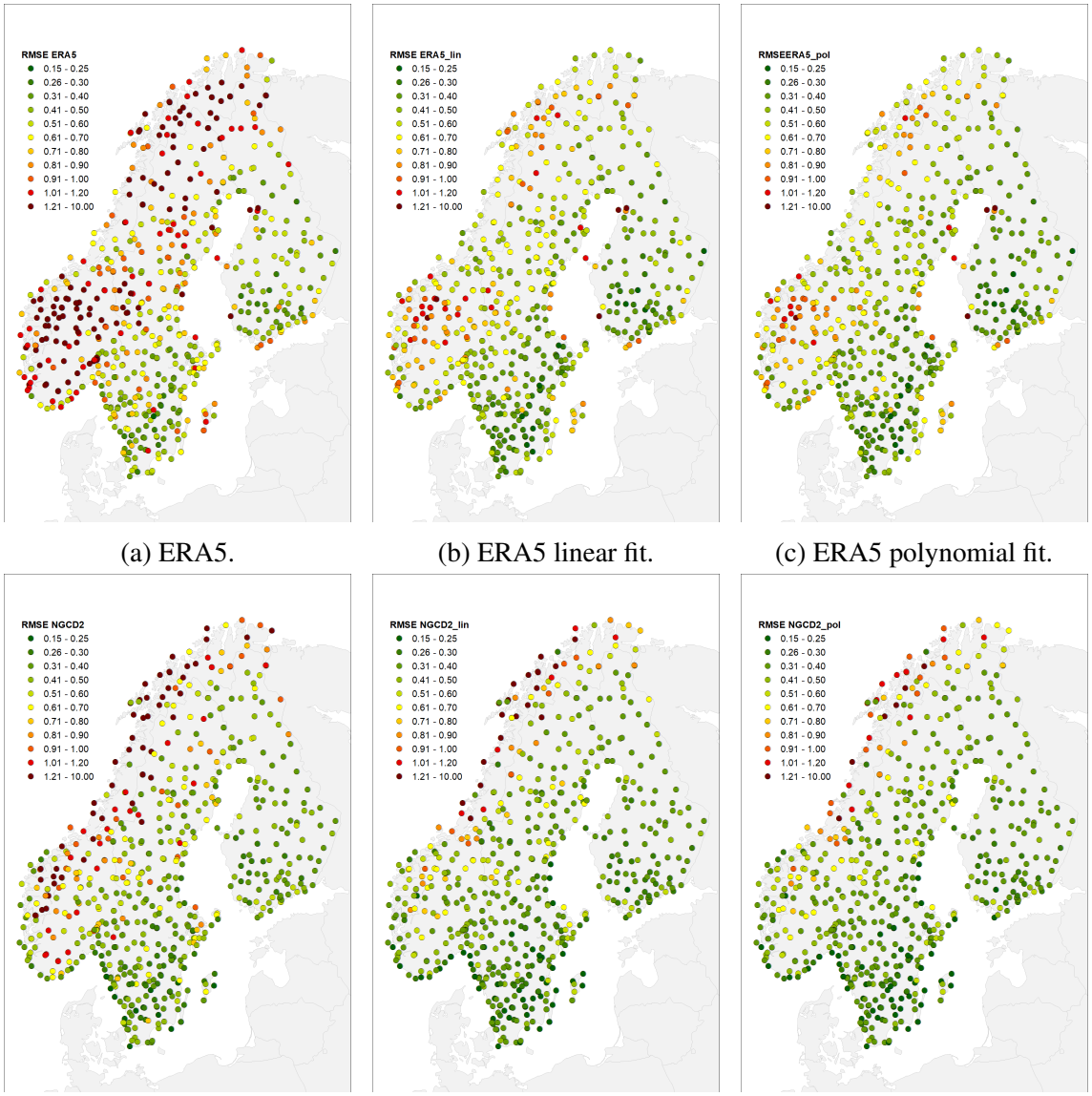


Figure 15: Distribution RMSE of downscaled temperatures from ERA5 and NGCD type 2. The black line shows the raw downscaled values while the red dashed line shows the bias-adjusted values using the linear adjustment described in equation 9. The blue dashed line represents the polynomial fit by equation 10.



(a) ERA5. (b) ERA5 linear fit. (c) ERA5 polynomial fit.

(d) NGCD2. (e) NGCD2 linear fit. (f) NGCD2 polynomial fit.

Figure 16: RMSE of downscaled datasets.

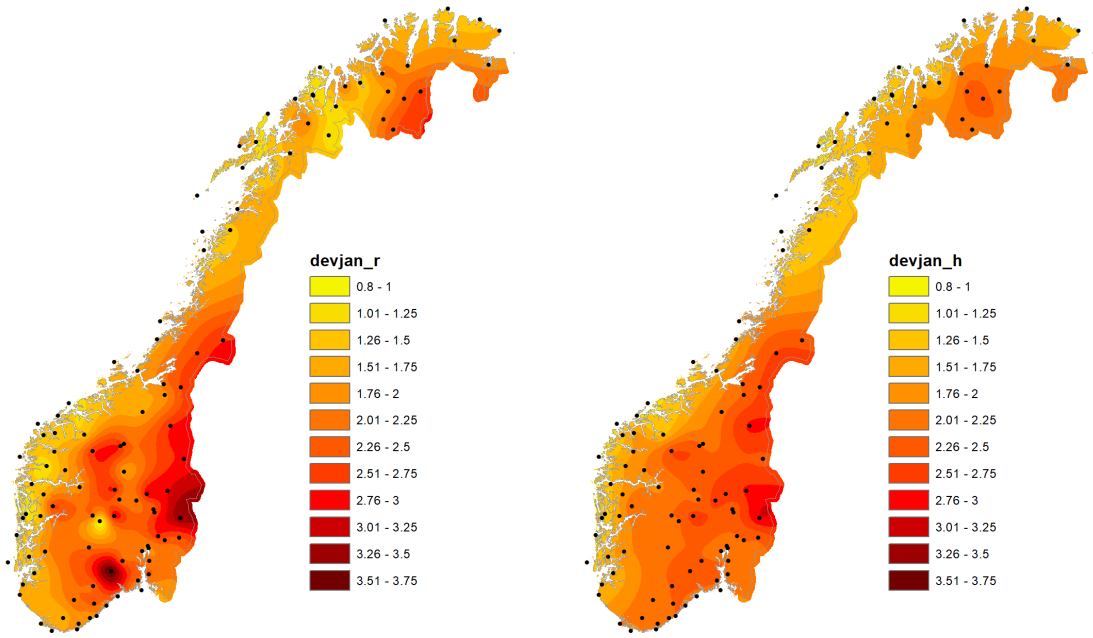
4 Some words on homogenisation

When analysing trends in observation series it is important that the series are homogeneous and only represent trends and changes in the weather and climate. Observation series can be severely influenced by external factors such as relocations, change of instruments, change of observers and/or observation practices and changes in the environment. The latter include buildings, land use and vegetation. These external factors might introduce an inhomogeneity in the series, either as an abrupt shift or as a gradual change. The first type of inhomogeneity is often caused by relocation, new equipment or new building nearby. Removal of vegetation close to the sensors is also a cause in this category. When vegetation is growing or the land-use is gradually changing the inhomogeneity will appear slowly, as a trend.

The first type of inhomogeneities are relatively easy to detect and adjust. There are a number of well tested methods developed for this purpose. The most applied methods for homogenisation in Europe are HOMER (*Mestre et al.*, 2013), ACMANT (*Domonkos*, 2019) and SNHT (*Alexandersson*, 1986). The SNHT method is implemented in the *Climatol* (*Guijarro*, 2018) R-package for analysing climate data.

The ClimNorm data set has not been completely homogeneity tested. The Norwegian series in the dataset is however tested and homogenized (*Kuya et al.*, 2020) applying HOMER. In the analysis series from Sweden and Finland were included to make the analysis of series near the national borders more robust. Only nine of the Norwegian series were found homogeneous throughout the 1961-1990 period. The most common reasons for inhomogeneities were relocations (43.8% of the breaks), automation (14.4%), new instruments (13%), new radiation screen (13%) and painting of radiation screen (9.1%).

The effect of the homogenisation can easy be illustrated by comparing maps of the change of monthly normal values between 1961-1990 and 1991-2020 based on "raw" unhomogenised series with maps based on homogenised series for January (Figure 17). The homogenised dataset shows a smooth change, representing regional climate trends, while the raw data result in a map with local variations. The maps based on homogenised data are clearly more trustworthy in explaining the large-scale climate variations that should explain the change of "normal" climatologies.



(a) Unhomogenized data.

(b) Homogenized data.

Figure 17: Difference between the 1961-1990 and 1991-2020 mean monthly January temperature.

5 Conclusions and recommendations

The ClimNorm project aims to support the national NMHS in the Nordic region in the calculation of new climate normals. For that purpose a joint climate dataset of monthly timeseries is established. This report has presented the temperature dataset, and methods to fill gaps. Two methods are actually applied and the results are compared. The preliminary investigations indicate that the method based on a principal component analysis tends to give more accurate estimates than applying bias-adjusted downscaling of gridded data. This is especially evident in areas where there are large local and regional variations in temperature due to elevation and coastal gradients. The PCA should be further tested by implementing and comparing the RSOI method (*Schiemann et al.*, 2010) and self organising map (SOM) algorithms that can estimate PCAs from series with missing data.

A homogenisation of the Norwegian and some Swedish and Finnish observation stations for the period 1961-2018 has shown that homogenised series shows more distinct regional patterns, and therefore can be regarded as better to describe climate variability and change than series that might be affected by artificial and environmental disturbances.

Further work should include:

- Update the ClimNorm data until 2019, and later also 2020.
- Test RSOI and SOM for filling gaps in time series
- Establish a gapfilled data set based on a recommended method.
- Homogenise the 1961-2020 data set.
- Analyse and homogenise longer time periods, e.g. 1901-2020.
- Study regional and temporal trends applying the gap-filled and homogenised data set.

References

Aalto, J., P. Pirinen, and K. Jylhä (2016), New gridded daily climatology of Finland: Permutation-based uncertainty estimates and temporal trends in climate, *Journal of Geophysical Research: Atmospheres*, *121*(8), 3807–3823.

Alexandersson, H. (1986), A homogeneity test applied to precipitation data, *Journal of Climatology*, *6*(6), 661–675, doi:10.1002/joc.3370060607.

C3S (2017), ERA5: Fifth generation of ECMWF atmospheric reanalyses of the global climate, *Tech. rep.*, Copernicus Climate Change Service.

Creutin, J. D., and C. Obled (1982), Objective analyses and mapping techniques for rainfall fields: An objective comparison, *Water Resources Research*, *18*(2), 413–431, doi:10.1029/WR018i002p00413.

Domonkos, P. (2019), The ACMANTv4 software package the ACMANTv4 software package, *Tech. rep.*, Available at: https://github.com/dpeterfree/ACMANT.

Eliassen, A. (1954), Provisional report on calculation of spatial covariance and autocorrelation of the pressure field, *Inst. Weather and Clim. Res., Acad. Sci., Oslo, Tech. Rep*, *5*.

Engeset, R., O. E. Tveito, H.-C. Udnæs, E. Alfnes, Z. Mengistu, K. Isaksen, and E. J. Førland (2004), Snow map validation for Norway, in *Proceedings XXIII Nordic Hydrological Conference 2004*, pp. 8–12.

Gandin, L. S., and R. Hardin (1965), *Objective analysis of meteorological fields*, vol. 242, Israel program for scientific translations Jerusalem.

Guijarro, J. (2018), Homogenizaton of climatic timeseries with climatol, version 3.1.1, *Tech. rep.*, doi:10.13140/RG.2.2.27020.41604.

Hersbach, H., B. Bell, P. Berrisford, S. Hirahara, A. Horányi, J. Muñoz-Sabater, J. Nicolas, C. Peubey, R. Radu, D. Schepers, A. Simmons, C. Soci, S. Abdalla, X. Abellan, G. Balsamo, P. Bechtold, G. Biavati, J. Bidlot, M. Bonavita, G. De Chiara, P. Dahlgren, D. Dee, M. Diamantakis, R. Dragani, J. Flemming, R. Forbes, M. Fuentes, A. Geer,

L. Haimberger, S. Healy, R. J. Hogan, E. Hólm, M. Janisková, S. Keeley, P. Laloyaux, P. Lopez, C. Lupu, G. Radnoti, P. de Rosnay, I. Rozum, F. Vamborg, S. Villaume, and J.-N. Thépaut (2020), The ERA5 global reanalysis, *Quarterly Journal of the Royal Meteorological Society*, *n/a*(n/a), doi:10.1002/qj.3803.

Hisdal, H., and O. E. Tveito (1992), Generation of runoff series at ungauged locations using empirical orthogonal functions in combination with kriging, *Stochastic Hydrology and Hydraulics*, *6*(4), 255–269.

Hisdal, H., and O. E. Tveito (1993), Extension of runoff series using empirical orthogonal functions, *Hydrological Sciences Journal*, *38*(1), 33–49, doi:10.1080/02626669309492638.

Kuya, E. K., H. M. Gjelten, and O. E. Tveito (2020), Homogenization of Norway's mean monthly temperature series.

Lussana, C., T. Saloranta, T. Skaugen, J. Magnusson, O. E. Tveito, and J. Andersen (2018a), seNorge2 daily precipitation, an observational gridded dataset over Norway from 1957 to the present day, *Earth System Science Data*, *10*(1), 235–249, doi:10.5194/essd-10-235-2018.

Lussana, C., O. E. Tveito, and F. Uboldi (2018b), Three-dimensional spatial interpolation of 2m temperature over Norway, *Quarterly Journal of the Royal Meteorological Society*, *144*(711), 344–364, doi:10.1002/qj.3208.

Mestre, O., P. Domonkos, F. Picard, I. Auer, S. Robin, E. Lebarbier, R. Böhm, E. Aguilar, J. A. Guijarro, G. Vertacnik, et al. (2013), HOMER: a homogenization software–methods and applications.

Schiemann, R., M. Liniger, and C. Frei (2010), Reduced space optimal interpolation of daily rain gauge precipitation in Switzerland, *Journal of Geophysical Research: Atmospheres*, *115*(D14).

Tveito, O., E. Førland, R. Heino, I. Hanssen-Bauer, H. Alexandersson, B. Dahlström, A. Drebs, C. Kern-Hansen, T. Jónsson, E. Vaarby Laursen, et al. (2000), Nordic temperature maps, *DNMI report*, *9*(00).

Tveito, O. E. (1998), Spatial estimation of mean monthly temperatures by multiple linear regression, *DNMI-klima*, *18*(98), 1–19.

Tveito, O. E., I. Bjørdal, A. O. Skjelvåg, and B. Aune (2005), A GIS-based agro-ecological decision system based on gridded climatology, *Meteorological Applications*, *12*(1), 57–68.

Wang, P., and M. Scharling (2010), Klimagrid Danmark - Dokumentation og validering af Klimagrid Danmark i 1x1 km opløsning.

WMO (2017), WMO Guidelines on the Calculation of Climate Normals, *Tech. Rep. WMO-No. 1203*, World Meteorological Organization.

WMO (2018), Guide to Climatological Practices, *Tech. Rep. WMO-No. 100*, World Meteorological Organization.