https://doi.org/10.5194/essd-2021-339
Preprint. Discussion started: 29 November 2021

EMO-5: A high-resolution multi-variable gridded meteorological data set for Europe

Vera Thiemig[1], Goncalo N. Gomes[1], Jon O. Skøien[1], Markus Ziese[2], Armin Rauthe-Schöch[2], Elke Rustemeier[2], Kira Rehfeldt[2], Jakub P. Walawender[2,5], Christine Kolbe[2,5], Damien Pichon[3], Christoph Schweim[4], Peter Salamon[1]

[1]European Commission, Joint Research Centre, Ispra, 21027, Italy
[2]Global Precipitation Climatology Centre, Deutscher Wetterdienst, Offenbach, 63067, Germany
[3]Kisters France SAS, Rueil-Malmaison, 92500, France
[4]Kisters AG, Aachen, 52076, Germany
[5]Faculty of Geography, Philipps University of Marburg, Marburg, 35032, Germany

Correspondence to: Peter Salamon (peter.salamon@ec.europa.eu)

Abstract. In this paper we present EMO-5[1], a European high-resolution, (sub-)daily, multi-variable meteorological data set built on historical and real-time observations obtained by integrating data from 18,964 ground weather stations, four high-resolution regional observational grids (i.e. CombiPrecip, ZAMG - INCA, EURO4M-APGD and CarpatClim) as well as one global reanalysis (ERA-Interim/Land). EMO-5 includes at daily resolution: total precipitation, temperatures (mean, minimum and maximum), wind speed, solar radiation and water vapour pressure. In addition, EMO-5 also makes available 6-hourly precipitation and mean temperature. The raw observations from the ground weather stations underwent a set of quality controls, before SPHEREMAP and Yamamoto interpolation methods were applied in order to estimate for each 5x5 km grid cell the variable value and its affiliated uncertainty, respectively. The quality of the EMO-5 precipitation data was evaluated through (1) comparison with two regional high resolution data sets (i.e. seNorge2 and seNorge2018), (2) analysis of 15 heavy precipitation events, and (3) examination of the interpolation uncertainty. Results show that EMO-5 successfully captured 80% of the heavy precipitation events, and that it is of comparable quality to a regional high resolution data set. The availability of the uncertainty fields increases the transparency of the data set and hence the possible usage. EMO-5 (release 1) covers the time period from 1990 to 2019, with a near real-time release of the latest gridded observations foreseen soon. As a product of Copernicus, the EU's Earth observation programme, EMO-5 dataset is free and open, and can be accessed at https://doi.org/10.2905/0BD84BE4-CEC8-4180-97A6-8B3ADAAC4D26 (Thiemig et al., 2021).

[1] EMO stands for "European Meteorological Observations", whereas the 5 denotes the spatial resolution of 5 km.

https://doi.org/10.5194/essd-2021-339
Preprint. Discussion started: 29 November 2021

Open Access Earth System Science Data Discussions

1 Introduction

The availability of meteorological data and their quality determine in many cases the capacity of environmental modelling. Easy accessibility to quality-controlled data, with good coverage and spatial resolution, can provide a solid foundation for various environmental modelling applications. For example, the European Flood Awareness System (EFAS), which is part of

the Emergency Management Service (EMS) of Copernicus, the EU's Earth observation programme, provides a flood monitoring and forecast service for riverine and flash floods across the whole of Europe. The forecasts of EFAS are calculated using the semi-distributed hydrological rainfall-runoff model LISFLOOD (https://ec-jrc.github.io/lisflood/) which relies heavily on quality-controlled, (sub-)daily meteorological information on precipitation, temperature, wind speed, solar radiation and water vapour pressure. Despite the existence of a fairly good coverage network of stations in Europe, there is currently no quality-controlled, high-resolution gridded meteorological data set available including all of the various meteorological variables, and the requirements of EFAS for near real-time and historical data.

For this reason, the Copernicus EMS Meteorological Data Collection Centre (MDCC) was created and tasked with the collection, quality control and gridding of real-time and historical meteorological data across Europe. For this release of the data set (not near real-time), 21 data providers contributed to the data collection, by sharing data from a total of about 18,694 in situ stations across Europe. It should be noted that while some stations measure multiple variables, others measure only one or two. Furthermore, some stations provided data for the entire period from 1970, while others only for a limited time-period. The MDCC collects 13 meteorological variables, all at the highest available temporal resolution and from 1970 onwards. In addition to the in situ stations, five gridded data sets have been added to the data collection, in order to improve the quality of the resulting meteorological grids by increasing the information density over those areas. Thus, daily 5x5 km grids have been created for seven variables - **precipitation**, **minimum, maximum, and mean air temperature**, **wind speed**, **solar radiation** and **water vapour pressure** - and additional 6-hourly grids for precipitation and mean air temperature. We refer to this data set as EMO-5[1].

At the present time, in addition to EMO-5 there are many other observational meteorological data grids, such as: E-OBS (Haylock et al., 2008; Cornes et al., 2018) for the whole of Europe; seNorge2 and seNorge2018 for Norway (Lussana et al., 2018); SPREAD for Spain (Serrano-Notivoli et al., 2017); ZAMG-INCA for Austria (Haiden et al., 2011); CombiPrecip for Switzerland (Sideris et al., 2014); CarpatClim for Hungary, Serbia, Romania, Ukraine, Slovakia, Poland, Czech Republic and Croatia (Antolović et al., 2013; Spinoni et al., 2015); EURO4M-APGD for the European Alps and adjacent flatland regions

https://doi.org/10.5194/essd-2021-339
Preprint. Discussion started: 29 November 2021

(Isotta et al. 2014); SAFRAN for France, Spain and Tunisia (Quintana-Seguí et al., 2008; Vidal et al., 2010; Quintana-Seguí et al., 2017; Tramblay et al., 2019). However, despite the availability of these data sets, EMO-5 represents a uniquely valuable resource due to a combination of its pan-European coverage, near real-time production, high temporal (6-hourly and daily) and spatial (5x5 km) resolution, large amount of input data (18,694 in situ stations and five high resolution regional observational grids), seven different variables (i.e. precipitation, minimum, maximum, and mean air temperature, wind speed, solar radiation and water vapour pressure), and a long historical data record (from 1 January 1990). These characteristics combine to make EMO-5, to our knowledge, the most complete gridded multi-variable observational meteorological data set covering the whole of Europe (and peripheral areas).

The aim of this paper is to present EMO-5 and its potential usage, by providing an insight into its data sources, the applied methods and the quality of the resulting information. The evaluation of the resulting grid quality focuses on the gridded precipitation data, as these are the most crucial drivers for environmental or hydrological modelling. The data set is already being successfully applied not only by EFAS, but also by two other major services of Copernicus EMS, namely the European Forest Fire Information System (EFFIS; https://effis.jrc.ec.europa.eu/; San-Miguel, J. et al., 2019) and the European Drought Observatory (EDO; https://edo.jrc.ec.europa.eu/; Spinoni et al., 2016; Cammalleri, C. et al., 2020). By making the EMO-5 data publicly available, we aim to support many other environmental applications and services that would benefit from using those data.

The remainder of this paper is organised as outlined in the following. The source data are described in Section 2. The entire workflow of the grid creation, including the quality control criteria applied during the data collection and an evaluation of various interpolation methods, are described in Section 3. Data access information are given in Section 4. An evaluation of the grid quality for precipitation is described in Section 5, and finally, some conclusions are presented in Section 6, followed by a future outlook.

2 Input data

The meteorological data for EMO-5 come from 26 data providers (i.e. 21 station data providers, plus 5 gridded data set providers), mostly being national meteorological services, and a few international or regional bodies (see Appendix 1 for the full list of data providers).

The MDCC collects historical and real-time observations obtained from 18,694 in situ meteorological stations across Europe, and an additional 13,394 meteorological stations globally. The MDCC collects 13 different meteorological variables: precipitation, 2m air temperature (i.e. measured at 2 metres above ground), daily minimum and maximum 2m air temperature, 10m wind speed (i.e. measured at 10 metres above ground), 10m wind direction, cloud cover, water vapour pressure, solar

https://doi.org/10.5194/essd-2021-339
Preprint. Discussion started: 29 November 2021

radiation, sunshine duration, relative air humidity, evaporation, and dew point temperature. Of these 13 variables, the following seven are used for EMO-5:

- Precipitation;
- Minimum, maximum and mean air temperatures;
- Wind speed;
- Water vapour pressure;
- Solar radiation.

Besides the actual in situ meteorological data, the station metadata including latitude, longitude, elevation, and (where available) instrument specifications, are collected. Data collection times depend on the data type (real-time or historical), the data availability and the chosen data transfer method. In general, real-time data are collected in a 24/7 mode as soon as they become available, whereas historical data are collected mostly on an annual basis. Regarding temporal resolution, for all variables except precipitation, the highest resolution (from 10 minutes upwards) is preferred as it provides the possibility to aggregate the data to multiple levels, which is useful both in its own right and for the quality control, as it increases its robustness. For precipitation on the other hand, the longest reported totals are used to calculate the daily and 6-hourly totals. Naturally, the number of variables per station varies, as does also the completeness of the data record per station.

Ingesting information from high-resolution regional observational grids has been shown to improve the quality of the final grids, in particular over areas with complex topography and / or low station density (Gampe and Ludwig, 2017). For this reason, the MDCC collects also data from four high-resolution regional gridded data sets (namely CombiPrecip, ZAMG-INCA, EURO4M-APGD and CarpatClim) as well as ERA-Interim/Land over areas with low station densities or with complex topography (see Fig. 1). For all regional observational gridded data sets having a spatial resolution higher than the one currently used in the European Flood Awareness System (implying all but CarpatClim), a regular subset of grid points with horizontal resolution of around 10x10 km was selected for integration into the MDCC data collection. The CarpatClim and ERA-Interim/Land data sets were imported at their original resolution. Each selected grid point is treated as a virtual station in the database. In total 10,632 virtual stations were added to the database for EMO-5. The main characteristics of the five input meteorological data grids underlying EMO-5 are summarized in Table 1:
[insert Table 1 here]

For each of the seven EMO-5 meteorological variables, the location (and hence density) of the input data, as well as the record length per station and the number of input stations over time (1990-2019), are shown in Fig. 1 to 3 respectively. As can be expected, the number of available stations for each variable and grid realisation is not constant over time or space. Jumps in the data coverage are caused by the integration of historical gridded data sets with fixed start and (partially) end dates, as well

https://doi.org/10.5194/essd-2021-339
Preprint. Discussion started: 29 November 2021

as the integration of historical data from data providers, beginning after 1990. This temporally and spatially inhomogeneous availability of stations leads to an inhomogeneous time series of grid-cell values, and therefore the EMO-5 data set is not optimised for trend (or temporal) analyses, but ideally suited as input data for a wide range of environmental and hydrological model applications.

[insert Figure 1, Figure 2 and Figure 3 here]

3 Methodology

3.1 Quality control on input data

All of the data collected by the MDCC undergo an automatic quality control procedure, irrespective of whether or not they have already been checked by their specific data provider. The quality control is implemented based on five types of data validation rules:

1) Availability: Check if value is present and time-stamp correct.
2) Monthly statistics: Check each value against statistical monthly data.
3) Cross-validation: Check each value against values from other parameters.
4) Minimum / maximum validation: Check each value against minimum / maximum thresholds.
5) Rate of change validation: Check the rate of change between two values against maximum thresholds.

The exact specification of the validation rules depends on the variable type as well as on the aggregation level, as summarized in Table 2.

[insert Table 2 here]

A data value is flagged as “missing” if the value is not available, and as “suspect” if the time stamp has been corrected (Rule 1). It is also flagged as “suspect” if it fails the validation against the monthly statistics (rule 2) or the cross-validation (rule 3). A data value is flagged as “rejected” if it falls outside of the defined minimum / maximum range (rule 4) or if the threshold for the maximum rate of change between two values has been exceeded (rule 5).

3.2 Choosing an optimal spatial interpolation scheme

As the primary usage of EMO-5 is to support an operational flood forecasting service, the data need to be delivered in a timely manner and at a good level of quality. Three spatial interpolation schemes - Inverse Distance Weighting (IDW), Modified SPHEREMAP, and Ordinary Kriging - were compared and evaluated in terms of reliability, specifically regarding uncertainty and computational cost. The three interpolation schemes are briefly described in Table 3.

https://doi.org/10.5194/essd-2021-339
Preprint. Discussion started: 29 November 2021

[insert Table 3 here]

The quality of each of the interpolation schemes was derived through a leave-one-out cross-validation. This means that for each iteration of the interpolated field, one station was left out and then later on compared with its interpolated value. This was done for around 4000 stations, and those pairs of interpolated and real observations were used to compute the uncertainty estimates. A similar approach was applied by Hofstra et al. (2008) for the E-CAD data set.

Three different measures of errors were calculated, which are the Mean Error (ME), Mean Absolute Error (MAE) and Mean Squared Error (MSE), as each focus on different aspects of uncertainties.

[insert Table 4 here]

As is shown in Table 4, SPHEREMAP was found to be the best scheme regarding ME, but IDW outperforms both other schemes regarding MAE and MSE. The variance between observed and computed value (MSE) is lowest for IDW for four out of six parameters. Only Kriging has lower variances for precipitation and vapour pressure. In addition, IDW has the lowest values for four out of six parameters regarding MAE. Here, SPHEREMAP performs a better interpolation for precipitation and vapour pressure. Regarding computation times, Kriging needed on average around 700 seconds, SPHEREMAP around 550 and IDW around 470. It should be noted that oceans were not masked to speed up the computations. Hence, SPHEREMAP was chosen as the interpolation scheme to generate the grids of EMO-5, as it shows the best performance for the critical parameter precipitation. For the other parameters, none of the tested schemes performed much better than the others. .

There are many methods available to estimate the uncertainty (i.e. reliability) of the gridded values, such as the leave-one-out approach, ensemble creations or the technique developed by Yamamoto (2000). Kriging itself provides an error estimation, but this depends only on the spatial distribution of the applied stations and not on the input data, therefore this error estimation is not applicable here. As the computational time of the grids is highly relevant in order to produce the operational grids as input for emergency management applications, the technique developed by Yamamoto (2000) is used due to its low computational effort. Furthermore, this method takes into account the variability of the surrounding observations, unlike the common Kriging uncertainty that only depends on the variogram and the spatial distribution of stations. This approach was also used, for example, for the E-OBS data set (Haylock et al., 2008). Originally, it was developed for Kriging schemes, but was adapted to the utilized modified SPHEREMAP scheme. Briefly, the method uses the interpolation weights to calculate a weighted variance between the gridded value and the input station data. It is zero if all input data are identical (e.g. areas with zero precipitation) and increases with increasing variance of the input data.

https://doi.org/10.5194/essd-2021-339
Preprint. Discussion started: 29 November 2021

Uncertainty fields were calculated and analysed for each grid realisation and interpolation method. Figure 4 below shows the gridded daily precipitation data and the corresponding uncertainty fields (for 15th May 2014) as an example for IDW (left), SPHEREMAP (middle) and Kriging (right).

[insert Figure 4 here]

The overall patterns for precipitation look very similar for the three interpolation approaches (Fig. 4, top row), with the only main difference being that Kriging produces smoother patterns compared with IDW and SPHEREMAP. Comparing their affiliated uncertainty patterns (Fig. 4, bottom row), SPHEREMAP shows the lowest uncertainty, and unlike the other two interpolation schemes the uncertainty signal of SPHEREMAP is geographically constrained to the regions with precipitation. Kriging shows extremely high gridding uncertainties, which we assume are due to a not well-fitted variogram.

The specific date (15th May 2014) is just one example that we selected from the comparison study of the interpolation methods, but it could have been any other day since the results were generally the same, to wit: 1) Kriging generated smoother variable fields than SPHEREMAP and IDW; and 2) SPHEREMAP showed overall the lowest and Kriging the largest uncertainty. Further, the interpolation performance (meaning the stability of the algorithm in a near real-time setting) was considered, as the timely availability of the gridded meteorological data must be assured in order to guarantee the smooth operation of the flood forecasting system. Unlike for IDW or SPHEREMAP, this is an issue for Kriging as the latter would require an automatic fitting of the variogram, which might go wrong in a near real-time operation. Based on the results of the uncertainty analysis, and the consideration of the interpolation performance, SPHEREMAP was chosen as the interpolation algorithm for EMO-5.

3.3 Grid creation

Moving from a large collection of in situ measurements to a European high-resolution, (sub-)daily, multi-variable meteorological data set, involves several processes and decisions, which are described in the following subsections.

3.3.1 Selection of stations

The number of stations used during gridding varies per variable and time step. This is due to the fact that for each grid creation the stations in the EFAS meteorological database are filtered based on a number of criteria. A station passes the filtering process for a particular variable and time step, if it fulfils all of the following criteria, in order to be used as input for gridding:

- The data coverage of aggregated precipitation readings is 100% for the entire period of the time step .
- The data coverage for other variables such as temperature, wind speed, etc. is 95% for daily data and 83% for 6-hourly mean temperature.
- The recorded values for that time-stamp passed the data quality check (see Section 3.1), meaning that they were flagged as “good” or “suspect” (data with quality “rejected” are excluded from gridding).

https://doi.org/10.5194/essd-2021-339
Preprint. Discussion started: 29 November 2021

For precipitation, not all stations that fulfil the above criteria are used in the interpolation. This is due to the fact that over time, there was an increasing number of identical stations that were reported by different data providers (e.g. identical stations are often found between the SYNOP as well as national data), albeit sometimes with slightly different values or slightly different coordinates. Not removing those duplicate stations would lead to a multiple counting of the same station during the interpolation, with the result of overweighting of those stations in the grids and less reliable area mean grid-cell values. To correct this, and to assure a gradual change between stations during the interpolation, duplicate stations within a vicinity of 500 metres were identified and merged into one virtual station. The coordinates of the virtual station were taken from the first station of this cluster, while the value was computed as the average of all duplicate stations. This reduced the total number of stations used per grid realisation by an average of 3.4%. Figure 1 shows the number of stations used during the grid creation.

Another filter that was implemented for precipitation on the level of the interpolation was a distance filter for ERA-Interim/Land data. The reason for this is that ERA-Interim/Land was used with the intention to fill the spatial gaps where no observations were available. However, with time the station density in those areas increased and therefore the need to integrate ERA-Interim/Land stations decreased. For this reason, all ERA-Interim/Land values that were less than 100 km away from a valid in situ station measurement were disregarded. Figure 1 shows how many of the 1402 existing virtual ERA-Interim stations were actually used for gridding.

3.3.2 Aggregation or reference period

Figure 5 shows the aggregation (i.e. reference period) for the different variables. Daily precipitation is accumulated from the previous day at 06:01 until the current day at 6:00. (Note: “current day” = “reference date of grid”). Daily mean temperature is calculated based on values observed between 18:01 on the previous day to 18:00 on the current day (see Section 3.3.5). The minimum temperature measured between 18:01 on the previous day until 06:00 on the current day is taken as the minimum temperature of the current day. On the other hand, the maximum measured temperature between 06:01 and 18:00 is used as the maximum temperature of the current day. All the other daily parameters such as wind speed, solar radiation and vapour pressure are averaged over 00:00 to 23:59:59 of the current day. For 6-hourly temperature, this depends on the data provider: if only 6-hourly instantaneous temperature readings are delivered, then those are used, otherwise temperature is averaged for 00:01-06:00, 06:01-12:00, 12:01-18:00, and 18:01-24:00 (see Section 3.3.5). The same time intervals are used to aggregate 6-hourly precipitation reported at 06:00, 12:00, 18:00 and 24:00 respectively. Note that the time-stamp of the EMO-5 grids refers to the end of the reference period, meaning that, for instance, the daily precipitation of 22.02.2021 06:00 covers the time period between 21.02.2021 06:01 and 22.02.2021 06:00.

[insert Figure 5 here]

https://doi.org/10.5194/essd-2021-339
Preprint. Discussion started: 29 November 2021

3.3.3 Land-sea mask

A land-sea mask is used to exclude sea surfaces from the gridding procedure. This is done mainly for two reasons. Firstly and most importantly, the number of meteorological in situ stations existing offshore is very small and stations are not homogeneously distributed. Both of these factors limit the potential quality and hence feasibility of gridding meteorological observations over sea surfaces. Secondly, and almost equally importantly, EMO-5 originates from the need for near real-time information on observed meteorological conditions over land surface areas. By excluding sea surface areas from the interpolation, the grid production is significantly faster and hence the delivery time is shorter.

3.3.4 Implications of altitude for temperature and water vapour pressure

In areas with strong orographic changes, neighbouring stations are likely to be at various altitudes, and hence the representativeness of their measurement for neighbouring areas is limited depending on the altitude change in the surrounding area (due to adiabatic processes). If station values were blindly used during the interpolation, an error would be introduced. This can be easily avoided by considering elevation information while interpolating. As SPHEREMAP does not take any auxiliary information into account, all recorded temperature and water vapour pressure values are brought to sea-level before interpolation. This is done based on altitude information obtained from a 1x1 km Digital Elevation Model (DEM), through accounting for the adiabatic change of 0.006 Kelvin per height metre for temperature and 0.00025 hPa / metre for water vapour pressure. The interpolation runs at sea-level and afterwards temperature and water vapour pressure information are brought back to the mean elevation of the respective 5x5 km grid cell by taking the parameter-specific correction factors into account.

3.3.5 Mean temperature

Where six hourly temperature averages are available, these have been used both for the six hourly data set and for the daily averages, aggregating the four 6-hourly observations. However, for many stations we did not have sub-daily temperature observations, particularly for the data from the 1990s. There are also many stations in the data set that have frequent observations between 06:00 and 18:00, enough for the 6-hourly daytime, but not for the nighttime period. This is particularly the case for observational data received from the MARS Meteorological Database of the European Commission's Joint Research Centre (Toreti et al., 2019). If data from other providers are not available at any station suffering from this issue, we have estimated the missing 6-hourly temperature based on the daily minimum and maximum temperatures. This was done based on the method described as the "Sin(14R-1)" method in Chow and Levermore (2007) with some small modifications.

For all stations where minimum and maximum temperatures are available, we fitted sinusoidal curves after estimating at what time of day the extremes most likely occurred. Similar to Chow and Levermore (2007), we assumed that the minimum occur 1 hour before sunrise, a value we computed with the function "getSunlightTimes" of the "suncalc" package (Thieurmel and Elmarhraoui, 2019) in the R programming language. However, we restricted the earliest time of the minimum temperature to

https://doi.org/10.5194/essd-2021-339
Preprint. Discussion started: 29 November 2021

occur not before 04:00 UTC in summer, with a correction for longitude. The latest time for the minimum was set at 10:00 UTC in winter, also corrected for longitude. While the "Sin(14R-1)" method assumes that the maximum temperature occurs at 14:00, we let the peak time range between 14:00 and 16:00 UTC, where it will be closer to 16:00 when sunset is late.

As some of the stations with missing 6-hourly data have stations with high-resolution observations nearby, we only included the estimated values for stations where the distance to a high-resolution neighbour is at least 10 km. This value was chosen by analyzing the mean square error for simulated temperatures at stations where we have 6-hourly averages and a variogram analysis of the 6-hourly data series. Requiring a mean square error of around 5.5 roughly corresponds to a distance of about 10 km in the variogram. Beyond this distance, it is likely that the errors from simulation are smaller than the spatial interpolation error. The number of extra stations is highest for the period 1990-1995, with around 500-600 stations with estimated 6-hourly temperatures. This can be compared with the approximately 1000 stations with 6-hourly temperature, as seen in Fig. 3. The number is then reduced to approximately 50-150 stations for most of the remaining period, but increases above 200 again for 2019 (mainly because of stations in Iceland).

Figure 6 shows the available stations for three different years (i.e. 1990, 2005 and 2019), for the 00:00 and 12:00 observations. The red dots show where we have 6-hourly temperature data, whereas the blue dots show where we have been able to add simulated 6-hourly temperature data based on the minimum - maximum observations. We can see that the simulated data covers much of Europe in 1990, whereas it drops for the later years. For the most recent years, most of the extra stations are added in North-Eastern Russia, some around the Mediterranean, and a large number in Iceland, where we have access to many more stations with minimum - maximum observations than 6-hourly data. For 1990, we can also notice the rather big difference between the data from nighttime (top) and daytime (bottom), where we have 6-hourly data from many more stations during the day than during the night.

[insert Figure 6 here]

3.3.6 Historical batch creation versus near real-time creation

The amount of station data available for gridding differs if these are created in near real-time or once in a batch for a given historical time period, with more stations available for the latter. This is due to the latency in data availability from climatological stations, which do not report in real-time mode but are included in historical data deliveries. The earliest station data are available and transferred into the MDCC database about 0.5 hours after the observational time, whereas all near real-time reporting stations including corrections are acquired within 5 days after the observational time. To account for the incremental filling of the database in the near real-time grid production, at every grid production cycle, grids of day-6 (i.e. six days ago) to day-2 are being produced (where "today" is defined as "day-0") and overwrite those previously produced. In this way, the latest grids always contain the highest amount of information available at any given moment.

https://doi.org/10.5194/essd-2021-339
Preprint. Discussion started: 29 November 2021

This incremental gridding strategy is not necessary if one would like to generate grids for a historical period, as the gridding is executed once in a batch process considering all the information available in the database at that given moment. The recreation of grids is performed according to need. Once a year all grids of the previous year are reproduced to account for the delivery of historical data from the data providers. The recreation of the whole time series is much rarer, and only happens after larger changes, for example after the change of the interpolation algorithm or the resolution (spatial or temporal), or the integration of larger amounts of historical data.

The EMO-5 version which will be available on the JRC Data Catalogue has been produced for the whole length of the archive from 1990 till 2019, and it is foreseen to extend this dataset on a monthly base. In order to have a data set that is consistent over time, it has been reproduced in a batch process for the entire time period using the modified SPHEREMAP algorithm and also for both temporal resolutions (6-hourly and daily).

4 Data availability

As a product of Copernicus, the EU's Earth observation programme, EMO-5 is free and open, and can be accessed at https://doi.org/10.2905/0BD84BE4-CEC8-4180-97A6-8B3ADAAC4D26 (Thiemig et al., 2021).

5 Evaluation

The quality of the EMO-5 precipitation data is evaluated in this section through (1) comparison with two regional high resolution data sets (i.e. seNorge2 and seNorge2018), (2) analysis of 15 heavy precipitation events, and (3) examination of the interpolation uncertainty.

5.1 Interpolation Uncertainty

As interpolated grid-values represent a "best guess", it is of paramount importance to provide information on the reliability. EMO-5 therefore contains, for each variable and time step, two fields - the interpolated value and an interpolation uncertainty, created with the methods described in Section 3.2.

An example of the associated interpolation uncertainty for the interpolated precipitation totals for the extreme event number 3 (see Section 5.3 below) is shown in Fig. 7. The highest accumulated precipitation totals are around 300 mm and the estimated accumulated interpolation uncertainty is up to 100 mm. In the largest parts of the extreme event area, the estimated accumulated precipitation uncertainty is around 25 mm, which is less than 20% of the accumulated precipitation total. Obviously, the highest estimated uncertainties are not located at the highest totals, but rather at the slopes of the highest rainfall areas. It is worth mentioning that the estimated interpolation uncertainty in the precipitation-free areas is roughly zero.

https://doi.org/10.5194/essd-2021-339
Preprint. Discussion started: 29 November 2021

[insert Figure 7 here.]

5.2 Comparison against a regional high resolution grid

We compared EMO-5 with two Norwegian data sets in order to assess its quality. One is the data set "seNorge2" (Lussana et al., 2018) and the other is the subsequent data set "seNorge_2018" (https://doi.org/10.5281/zenodo.2082320, Lussana, 2018b). The area covered by these two data sets consists of the Norwegian mainland and a strip of the neighbouring countries Sweden, Finland, and Russia (Lussana, 2018).

The covered area is very challenging for raster data generation with large differences in the orography, strong precipitation and large small-scale precipitation differences (e.g. due to the orography). Both seNorge data sets are based on station data from the Norwegian Climate Database (data.met.no) and outside Norway on the European Climate Assessment data set (Klein Tank et al., 2002). For the seNorge2 data set the raw data are used while for the seNorge2018 data set an undercatch correction is applied, which has a large impact especially in the winter months when snow is detected. The data are on a regular grid with 1-km grid spacing for the periods 1957-2015 for seNorge2 and 1957-2017 for seNorge2018. Both data sets have daily values from 6 UTC to 6 UTC. An extra interpolation based on nearest neighbour is used to identify the precipitation days per grid point, or otherwise to set the grid values to zero. The interpolation procedure for the precipitation values for both data sets is based on an optimal interpolation procedure with a background field. The seNorge2 data set uses the station data as a background field, while seNorge2018 uses the monthly values of the ERA-Interim analysis with 2.5 degrees resolution. For the actual interpolation, a cascade of decreasing scales is used and the station height is taken into account. The differences are mainly seen in mountainous regions with sparse station coverage. Since the optimal interpolation procedure requires a normal distribution, a Box-Cox transformation is applied to the seNorge2018 data set (Lussana et al., 2019).

For comparison with EMO-5, the period 1990-2015 which is common to all data sets, is used. The day definition (from 6 UTC to 6 UTC) is the same for all three data sets. To obtain the same spatial projection, the seNorge data sets were reshaped to the 5-km resolution of EMO-5. For this purpose, the bilinear interpolation method was applied, which uses a weighted average of the 4 nearest grid points. The study area of this analysis can be seen in Fig. 8.

[insert Figure 8 here.]

The differences in annual precipitation, as well as the differences in the distributions of seasonal precipitation, and the distributions of extremes represented by the precipitation indices adopted from the CCl/WCRP/JCOMM Expert Team on Climate Change Detection and Indices (ETCCDI) set by Peterson et al. (2001, Appendix A), are presented below.

As can be seen in Fig. 8, the data have a similar spatial structure, with seNorge2018 showing heavier precipitation. This is more clearly visible in the direct comparison in Fig. 9.

[insert Figure 9 here.]

https://doi.org/10.5194/essd-2021-339
Preprint. Discussion started: 29 November 2021

Open Access Earth System Science Data Discussions

Figure 9 shows that EMO-5 and seNorge2 have very similar magnitudes of values. The differences between the two data sets are very small-scale, especially in the southwest and northeast. These could indicate orographic effects, as the seNorge interpolation methods include the elevation as a parameter. Large structures can be found in the centre of the area, which could indicate fundamental data differences. In comparison with seNorge2018, the high precipitation values of seNorge2 are evident again. This is not surprising due to the undercatch correction in this data set. However, two blue patches are conspicuous, showing that EMO-5 is characterized by higher precipitation than seNorge2018. The seasonal differences are listed in Table 5 (differences in the values shown in Table 5 and Fig. 9 are due to the slightly different procedure to derive the maximum).
[insert Table 5 here.]

The average annual mean precipitation of EMO-5 is 2% higher compared to the one of seNorge2, and 17% lower compared to seNorge2018. In the average maximum precipitation per grid, EMO-5 shows 106% of the seNorge2 precipitation and 86% of the seNorge2018 precipitation. EMO-5 is thus substantially closer to seNorge2 in mean grid precipitation than the two seNorge data sets are to each other (see Table 5).
[insert Figure 10 here.]

The comparison of the seasonal precipitation in Fig. 10(a) shows that the individual months are very different. There are very large differences especially with respect to intense precipitation. These are distinctly higher in EMO-5 than in the seNorge data sets, an exception being autumn. In this case, the distributions of seNorge2018 and EMO-5 agree very well. Large EMO-5 data values are rare, only 1523 values (or 0.06%) are above 1500mm, and 212 values (9^10-5%) are above 2000mm. The reason is probably data errors, as these values can be located in a few small locations in the study area. For the benefit of visual comparability, the highest 0.01% of the values are omitted in Fig. 10.

The probability distribution of the smaller values, on the other hand, shows very good agreement with seNorge2, while seNorge2018 shows larger values especially in spring and winter. The comparison of the two seNorge data sets shows significantly higher values for seNorge2018, especially in spring and winter. In summer, the seNorge data sets are very similar. The latter finding can be explained very well by the fact that the undercatch correction has less impact in summer. The results look more diverse when considering the extremes on a daily basis (see Fig. 10(b)). The maximum extreme value indices are calculated over the entire time period for each grid point. Again, as in Fig. 10(a), the highest 0.01% of the respective values are not shown. This concerns for example the high values of the Consecutive Dry Days (CDD) - up to 1400 days without precipitation - in the EMO-5 data set. These can be limited to a small region in the northeast. The high values for the Consecutive Wet Days (CWD) can be attributed to two points in the centre and one in the south of the study area. Otherwise, the extra interpolation for the Yes-No decision for precipitation days in the seNorge data sets could have an influence on the distributions of the CDD and CWD indices. The Simple Daily Intensity Index (SDII) shows similar distributions as the totals

https://doi.org/10.5194/essd-2021-339
Preprint. Discussion started: 29 November 2021

in Fig. 10(a). SeNorge2 and EMO-5 agree very well, while seNorge2018 shows much higher values. The number of precipitation days above 10 mm and above 20 mm (not shown) shows a good agreement between seNorge2 and EMO-5, similar to SDII, with a slightly better agreement at r10. The maximum 5-day totals show a strikingly different picture. Here EMO-5 agrees very well with seNorge2018, while seNorge2 shows significantly lower values.

In summary, the agreement between the seNorge2 and EMO-5 data sets is very good, despite the very different interpolation methods. The influence of the undercatch correction used in the seNorge_2018 data set is substantially larger. In particular, the lower precipitation values seem to match very well in seNorge2 and EMO-5. This is also evident in the extreme value indices based on low values, especially r10. The ratio of the annual values clearly shows the different representation of the orographic structure due to the orography-dependent interpolation in the seNorge data sets. The large differences in the extremes with very few values are mainly due to isolated data errors in the EMO-5 data. However, the differences in the isolated grid points do not appear to be caused by single extremely high daily values, as these would otherwise show in indices such as the Maximum Consecutive 5-Day Precipitation (RX5day). Instead, smaller but longer-lasting errors seem to cause these discrepancies. These then affect statistics that use all values over a period of time, such as the seasonal totals or indices like CWD and CDD.

5.3 Heavy precipitation events

As well as the timeliness of EMO-5, also the capability to capture heavy precipitation events is of great importance for EFAS. In this Section, we evaluate semi-quantitatively the ability of EMO-5 to capture heavy precipitation events. We first randomly selected from www.floodlist.com, 15 flood events that were caused by heavy precipitation. Affiliated precipitation information was then extracted and verified through visual inspection of the respective EMO-5 image.

Figure 11 shows for each of the 15 selected flood events the related precipitation information as published on Floodlist (on the left-hand side) and as seen in the EMO-5 data set (all four maps). The information regarding the observed precipitation as published on Floodlist ranges from a qualitative indication, such as "heavy precipitation between those dates" (Events 2, 3, 4, 11 and 14) to precise quantitative statements such as "x amount of precipitation between those dates" (Events 1, 5-10, 12, 13 and 15). When the original data source of the observed precipitation was mentioned on Floodlist, then it is also cited here. For all the other cases the value shall just be used as an indication.

Regarding EMO-5, we present here the sum of daily and 6-hourly precipitation maps covering the dates as specified on Floodlist, as well as the maximum daily and 6-hourly precipitation values within that time frame. In order to cover a full calendar day it was necessary to sum up 2 daily precipitation maps. This is due to the fact that the aggregation or reference period of the daily precipitation maps is not in line with a calendar day, but covers the period from 06:01 of the previous day until 06:00 of the current day (see Section 3.3.2). Hence, there is always one more daily map than the number of calendar days to cover. This was not necessary for the 6-hourly maps, as four 6-hourly maps cover precisely one calendar day. This explains

https://doi.org/10.5194/essd-2021-339
Preprint. Discussion started: 29 November 2021

also why in Fig. 11 the sum of daily precipitation is generally higher than the sum of the 6-hourly precipitation. This should not be a problem, as the aim of this comparison is to evaluate the performance of the daily and 6-hourly precipitation against the reported information, and not to compare the 6-hourly with the daily totals.

For 13 out of the 15 selected events, EMO-5 shows larger amounts of precipitation sums over the event duration ranging between 119 mm (Event 5) to 350 mm (Event 9), with 270 mm and 255 mm being the maximum observed precipitation within a daily and 6-hourly period respectively. The only event clearly missed is Event 6, where hardly any precipitation was captured by EMO-5. This event was caused by a cloud burst, and hence was a very local extreme event. It is likely that it was missed as no in situ station exists directly at the specific location in EMO-5.

For the nine events with concrete specifications on the precipitation amounts, EMO-5 captures the precipitation amounts of three events (Events 5, 8 and 15) in accordance with the media reports, overestimates the amount for one event (Event 9), and underestimates the reported amounts for five events (Events 6, 7, 10, 12 and 13) . Three out of the five underestimated precipitation events (Events 6, 10 and 12) were also caused by cloud bursts, and hence were short and high intensity events that are difficult to reproduce in a temporally aggregated and spatially interpolated data set, especially if the full amount of precipitation was not captured by any in situ station, due to the positioning of the station.
[insert Figure 11 here]

6 Conclusions and future work

EMO-5 is a European high-resolution (5x5 km), (sub-)daily, multi-variable, multi-decadal meteorological data set based on quality-controlled observations coming from almost 30,000 stations (18,964 in situ and 10,632 virtual stations) across Europe, and is produced in near real-time. The data set covers precipitation, temperature (mean, minimum and maximum), wind speed, solar radiation and water vapour pressure all at a daily resolution, and in addition 6-hourly for precipitation and mean temperature. EMO-5 covers the time period from 1990 onwards and is freely available. In this paper we have provided insight into the source data, the applied methods and the quality assessment of EMO-5 (the latter just for precipitation).

EMO-5 grids are produced by means of a modified SPHEREMAP algorithmus, which is a geometric scheme taken the distance, clustering and gradient into account. The decision was made after a comparison of three interpolation schemes. An algorithm was developed to identify stations provided independently by different data providers and replace them by a merged station.
Each EMO-5 grid realisation is accompanied by a corresponding estimate of the interpolation uncertainty, which is based on the approach by Yamamoto (2000). This takes the difference between the gridded value and the station data as well as the interpolation weights into account. Therefore, the estimated uncertainty at each grid cell differs from day to day. The highest

https://doi.org/10.5194/essd-2021-339
Preprint. Discussion started: 29 November 2021

uncertainties of the gridded data are at steep gradients between regions with high and low precipitation totals, reflecting the uncertainty in the estimation of the spatial extension of such an event.

The quality of the precipitation product was evaluated through comparison against a regional high-resolution data set (seNorge2 and seNorge2018) over Norway as well as against 15 media reports of extreme precipitation events that caused flooding across Europe.

The comparison of EMO-5 against two regional high resolution data sets over Norway has shown that EMO-5 and seNorge2 are more similar than both Norwegian data sets are to each other. The good agreement is despite their very different interpolation methods, and is particularly evident for lower precipitation values and mean annual means, while differences can be detected in the seasonality and the extreme values. The latter issue is suspected to be caused by longer-lasting data errors of very few isolated stations used in the EMO-5 data set. The reason for seNorge2018 being substantially different to both EMO-5 and seNorge2, is likely due to the undercatch correction that is applied only to seNorge2018. It is recommended to broaden this comparison to other high resolution data sets before drawing general conclusions. However, the comparison done here has shown that EMO-5 can be comparable in terms of quality with a regional high-resolution data set.

The semi-quantitative evaluation of EMO-5's capacity to capture heavy precipitation events has shown that in 80% of the cases, EMO-5 captures the events qualitatively, meaning that the EMO-5 grids covering the event show large amounts of precipitation, which is important in particular for applications such as flood modelling, that rely heavily on the forcing data to be able to capture those events. As expected for gridded data sets, even if EMO-5 mostly captures heavy precipitation events, it tends to underestimate the observed precipitation amount at stations. This is not a surprising finding as especially convective events are of short temporal and spatial scale. This makes it very difficult to capture the maximum precipitation amount of those events, unless an in situ station is in the direct vicinity of the event and captures it.

As part of the operational Copernicus EMS, the number of stations (historical and near real-time) that are used for gridding in EMO-5 will be continuously increased, through adding new data providers and the integration of new, high-resolution regional observational grids, where available. In addition, to improve the current quality control framework, new data validation rules, such as spatial comparison with neighbouring stations or additional statistical checks, will be implemented Finally, as it is foreseen to increase the spatial resolution of EFAS from the current 5 km grid to a 1 arc minute grid (approximately equal to 1.8 km at the equator), also EMO-5 will increase the spatial resolution in its next version.

Acknowledgements

Our sincere gratitude goes to all the data providers who shared and continue sharing their (station) data with the CEMS Meteorological Data Collection Centre (CEMS MDCC). A full list can be found in the supplementary material. The CEMS MDCC is funded by the Copernicus program, and runs as part of the Copernicus Emergency Management Service (CEMS).

We would furthermore like to thank the Copernicus in situ cross coordination action implemented by the European Environmental Agency for its support in setting up in situ data licenses for meteorological data; and lastly, Niall McCormick (JRC) for English proofreading and Christel Prudhomme (ECMWF), Carlo Buontempo (C3S) and Andrea Toreti (JRC MARS) for reviewing.

References

Antolović, I., Mihajlović, V., Rančić, D., Mihić, D. and Djurdjević, V.: Digital Climate Atlas of the Carpathian Region, Advances in Science and Research, 10(1), 107–111, https://doi.org/10.5194/asr-10-107-2013, 2013.

Cammalleri, C., Arias Muñoz, C., Barbosa, P., de Jager, A., Magni, D., Masante, D., Mazzeschi, M., Mccormick, N., Naumann, G., Spinoni, J. and Vogt, J.: A revision of the Combined Drought Indicator (CDI) as part of the European Drought Observatory (EDO)., 2020.

CCl/WCRP/JCOMM Expert Team on Climate Change Detection and Indices: Climate Change Indices - Definitions of the 27 core indices http://etccdi.pacificclimate.org/list_27_indices.shtml, last checked 8.3.2021

Chow, D. H. C. and Levermore, G. J.: New algorithm for generating hourly temperature values using daily maximum, minimum and average values from climate models, Building Services Engineering Research and Technology, 28(3), 237–248, https://doi.org/10.1177/0143624407078642, 2007.

Cornes, R. C., Schrier, G. van der, Besselaar, E. J. M. van den and Jones, P. D.: An Ensemble Version of the E-OBS Temperature and Precipitation Data Sets, Journal of Geophysical Research: Atmospheres, 123(17), 9391–9409, https://doi.org/10.1029/2017JD028200, 2018.

Dee, D.: ERA-Interim, ECMWF https://www.ecmwf.int/en/forecasts/data sets/reanalysis-data sets/era-interim, last access: 17 February 2021, 2011.

Dee, D. P., Uppala, S. M., Simmons, A. J., Berrisford, P., Poli, P., Kobayashi, S., Andrae, U., Balmaseda, M. A., Balsamo, G., Bauer, P., Bechtold, P., Beljaars, A. C. M., Berg, L. van de, Bidlot, J., Bormann, N., Delsol, C., Dragani, R., Fuentes, M., Geer, A. J., Haimberger, L., Healy, S. B., Hersbach, H., Hólm, E. V., Isaksen, L., Kållberg, P., Köhler, M., Matricardi, M., McNally, A. P., Monge-Sanz, B. M., Morcrette, J.-J., Park, B.-K., Peubey, C., Rosnay, P. de, Tavolato, C., Thépaut, J.-N. and Vitart, F.: The ERA-Interim reanalysis: configuration and performance of the data assimilation system, Quarterly Journal of the Royal Meteorological Society, 137(656), 553–597, https://doi.org/10.1002/qj.828, 2011.

Gampe, D. and Ludwig, R.: Evaluation of Gridded Precipitation Data Products for Hydrological Applications in Complex Topography, Hydrology, 4(4), https://doi.org/10.3390/hydrology4040053, 2017.

Haiden, T., Kann, A., Wittmann, C., Pistotnik, G., Bica, B. and Gruber, C.: The Integrated Nowcasting through Comprehensive Analysis (INCA) System and Its Validation over the Eastern Alpine Region, Weather and Forecasting, 26(2), 166–183, https://doi.org/10.1175/2010WAF2222451.1, 2011.

Haylock, M. R., Hofstra, N., Tank, A. M. G. K., Klok, E. J., Jones, P. D. and New, M.: A European daily high-resolution gridded data set of surface temperature and precipitation for 1950–2006, Journal of Geophysical Research: Atmospheres, 113(D20), https://doi.org/10.1029/2008JD010201, 2008.

Hofstra, N., Haylock, M., New, M., Jones, P. and Frei, C.: Comparison of six methods for the interpolation of daily, European climate data, Journal of Geophysical Research: Atmospheres, 113(D21), https://doi.org/10.1029/2008JD010100, 2008.

Isotta, F. A., Frei, C., Weilguni, V., Tadić, M. P., Lassègues, P., Rudolf, B., Pavan, V., Cacciamani, C., Antolini, G., Ratto, S. M., Munari, M., Micheletti, S., Bonati, V., Lussana, C., Ronchi, C., Panettieri, E., Marigo, G. and Vertačnik, G.: The climate of daily precipitation in the Alps: development and analysis of a high-resolution grid data set from pan-Alpine rain-gauge data, International Journal of Climatology, 34(5), 1657–1675, https://doi.org/10.1002/joc.3794, 2014.

Krige, D. G.: Two-dimensional weighted moving average trend surfaces for ore-evaluation, Journal of the South African Institute of Mining and Metallurgy, 66, 13–38, 1966.

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

Lussana, C.: seNorge_2018 daily total precipitation amount 1957–2017, https://doi.org/10.5281/zenodo.2082320, the Norwegian Meteorological Institute, 2018b.

Lussana, Cristian and Tveito, Ole Einar and Dobler, Andreas and Tunheim, Ketil: SeNorge 2018, daily precipitation, and temperature data sets over Norway, Earth System Science Data, 11(4), 1531–1551, 2019.

Klein Tank, A. M., Wijngaard, J. B., Können, G. P., Böhm, R., Demarée, G., Gocheva, A., Mileta, M., Pashiardis, S., Hejkrlik, L., Kern-Hansen, C., Heino, R., Bessemoulin, P., Müller-Westermeier, G., Tzanakou, M., Szalai, S., Pálsdóttir, T., Fitzgerald, D., Rubin, S., Capaldo, M., Maugeri, M., Leitass, A., Bukantis, A., Aberfeld, R., van Engelen, A. F., Forland, E., Mietus, M., Coelho, F., Mares, C., Razuvaev, V., Nieplova, E., Cegnar, T., Antonio López, J., Dahlström, B., Moberg, A., Kirchhofer, W., Ceylan, A., Pachaliuk, O., Alexander, L. V., and Petrovic, P. : Daily data set of 20th-century surface air temperature and precipitation series for the European Climate Assessment, Int. J. Climatol., 22, 1441–1453, 2002.

Peterson, T., Folland, C., Gruza, G., Hogg, W., Mokssit, A. and Plummer, N.: Report on the activities of the working group on climate change detection and related rapporteurs. Geneva: World Meteorological Organization, 2001.

Quintana-Seguí, P., Moigne, P. L., Durand, Y., Martin, E., Habets, F., Baillon, M., Canellas, C., Franchisteguy, L. and Morel, S.: Analysis of Near-Surface Atmospheric Variables: Validation of the SAFRAN Analysis over France, Journal of Applied Meteorology and Climatology, 47(1), 92–107, https://doi.org/10.1175/2007JAMC1636.1, 2008.

Quintana-Seguí, P., Peral, C., Turco, M., Llasat, M. C. and Martin, E.: Meteorological Analysis Systems in North-East Spain: Validation of SAFRAN and SPAN, JOURNAL OF ENVIRONMENTAL INFORMATICS, 27(2), 116-130–130, 2017.

San-Miguel-Ayanz, J., Durrant, T., Boca, R., Libertà, G., Branco, A., de Rigo, D., Ferrari, D., Maianti, P., Artés Vivancos, T., Oom, D., Pfeiffer, H., Nuijten, D., Leray, T. and Moffat, A.J.: Forest Fires in Europe, Middle East and North Africa 2018. EUR 29856 EN, ISBN 978-92-76-11234-1, https://doi.org/10.2760/1128, 2019.

Schamm, K., Ziese, M., Becker, A., Finger, P., Meyer-Christoffer, A., Schneider, U., Schröder, M. and Stender, P.: Global gridded precipitation over land: a description of the new GPCC First Guess Daily product, Earth System Science Data, 6(1), 49–60, https://doi.org/10.5194/essd-6-49-2014, 2014.

Serrano-Notivoli, R., Beguería, S., Saz, M. Á., Longares, L. A. and de Luis, M.: SPREAD: a high-resolution daily gridded precipitation data set for Spain – an extreme events frequency and intensity overview, Earth System Science Data, 9(2), 721–738, https://doi.org/10.5194/essd-9-721-2017, 2017.

Sideris, I. V., Gabella, M., Erdin, R. and Germann, U.: Real-time radar–rain-gauge merging using spatio-temporal co-kriging with external drift in the alpine terrain of Switzerland, Quarterly Journal of the Royal Meteorological Society, 140(680), 1097–1111, https://doi.org/10.1002/qj.2188, 2014.

Shepard, D. B.: A two-dimensional interpolation function for irregularly-spaced data, Proceedings of the 1968 23rd ACM national conference, https://doi.org/10.1145/800186.810616, 1968.

Spinoni, J., Naumann, G., Vogt, J. and Barbosa, P.: Meteorological Droughts in Europe: Events and Impacts - Past Trends and Future Projections, Publications Office of the European Union., https://doi.org/10.2788/450449, 2016.

Spinoni, J., Szalai, S., Szentimrey, T., Lakatos, M., Bihari, Z., Nagy, A., Németh, Á., Kovács, T., Mihic, D., Dacic, M., Petrovic, P., Kržič, A., Hiebl, J., Auer, I., Milkovic, J., Štepánek, P., Zahradnícek, P., Kilar, P., Limanowka, D., Pyrc, R., Cheval, S., Birsan, M.-V., Dumitrescu, A., Deak, G., Matei, M., Antolovic, I., Nejedlík, P., Štastný, P., Kajaba, P., Bochnícek, O., Galo, D., Mikulová, K., Nabyvanets, Y., Skrynyk, O., Krakovska, S., Gnatiuk, N., Tolasz, R., Antofie, T. and Vogt, J.: Climate of the Carpathian Region in the period 1961–2010: climatologies and trends of 10 variables, International Journal of Climatology, 35(7), 1322–1341, https://doi.org/10.1002/joc.4059, 2015.

Thieurmel, B. and Elmarhraoui, A.: R-package suncalc: Compute Sun Position, Sunlight Phases, Moon Position and Lunar Phase. https://CRAN.R-project.org/package=suncalc, last access: 12 February 2021, 2019.

Toreti, A., A. Maiorano, G. De Sanctis, H. Webber, A. C. Ruane, D. Fumagalli, A. Ceglar, S. Niemeyer, and M. Zampieri: Using Reanalysis in Crop Monitoring and Forecasting Systems. Agricultural Systems, 168: 144–53. https://doi.org/10.1016/j.agsy.2018.07.001, 2019.

Tramblay, Y., Feki, H., Quintana-Seguí, P. and Guijarro, J. A.: The SAFRAN daily gridded precipitation product in Tunisia (1979–2015), International Journal of Climatology, 39(15), 5830–5838, https://doi.org/10.1002/joc.6181, 2019.

Vidal, J.-P., Martin, E., Franchistéguy, L., Baillon, M. and Soubeyroux, J.-M.: A 50-year high-resolution atmospheric reanalysis over France with the Safran system, International Journal of Climatology, 30(11), 1627–1644, https://doi.org/10.1002/joc.2003, 2010.

Willmott, C., Rowe, C. and Philpot, W.: Small-Scale Climate Maps: A Sensitivity Analysis of Some Common Assumptions Associated with Grid-Point Interpolation and Contouring, American Cartographer, 12, 5–16, https://doi.org/10.1559/152304085783914686, 1985.

Willmott, C. J. and Matsuura, K.: Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance, Climate Research, 30(1), 79–82, https://doi.org/10.3354/cr030079, 2005.

https://doi.org/10.5194/essd-2021-339
Preprint. Discussion started: 29 November 2021

Earth System Science Data
Open Access
Discussions

Yamamoto, J. K.: An Alternative Measure of the Reliability of Ordinary Kriging Estimates, Mathematical Geology, 32(4), 489–509, https://doi.org/10.1023/A:1007577916868, 2000.

Figures

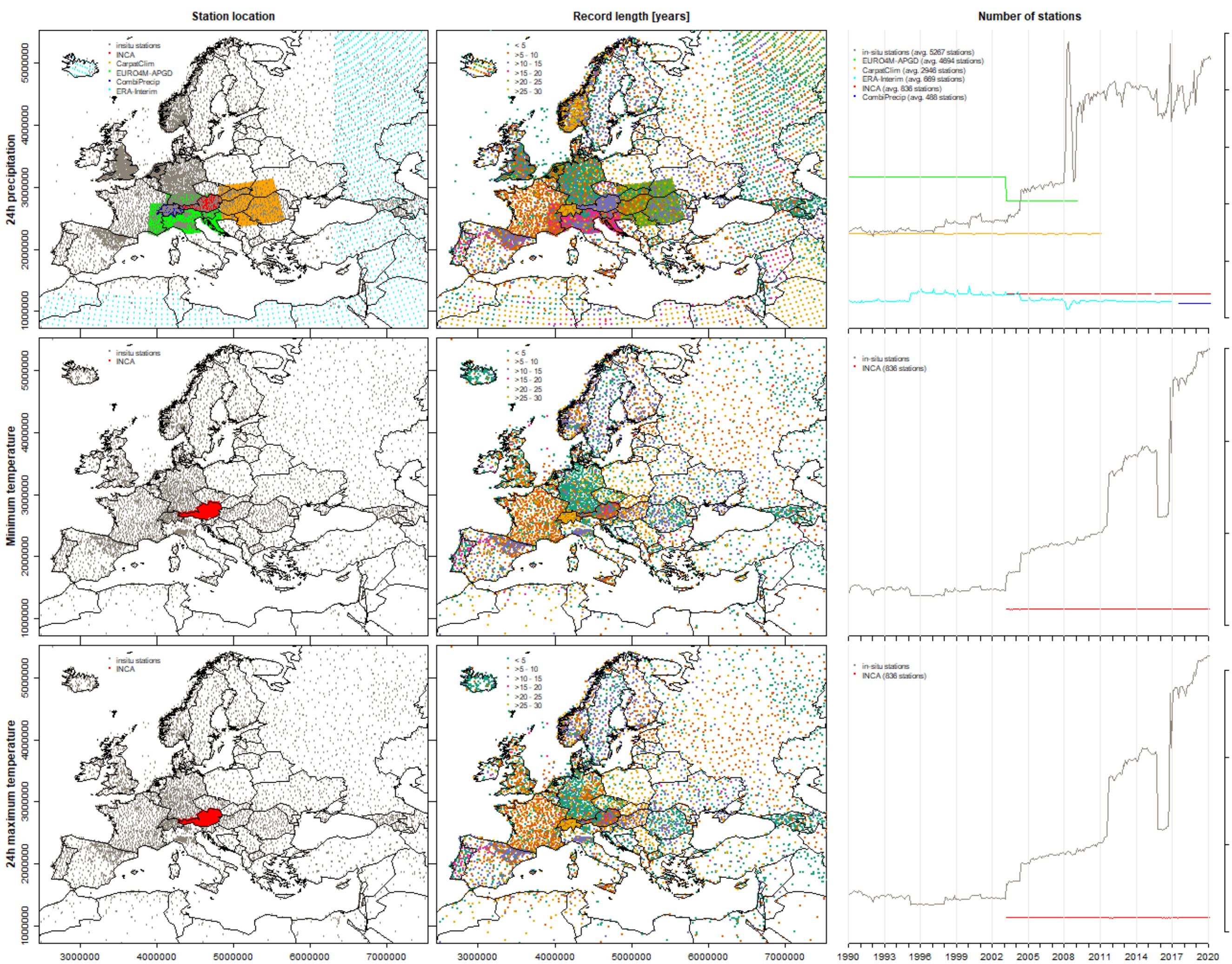
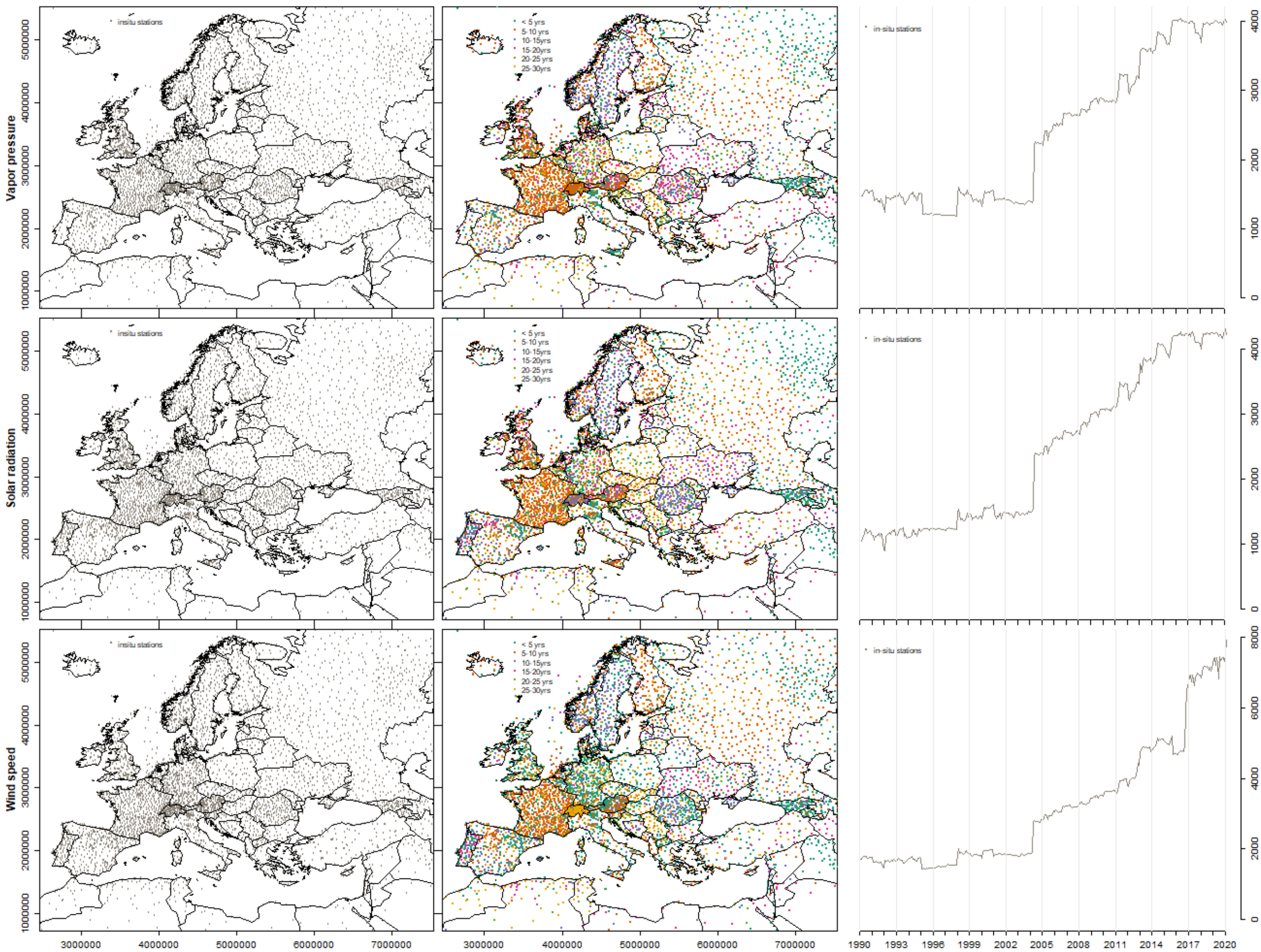


Figure 1: Input data for daily EMO-5 precipitation and minimum and maximum temperature grids in terms of station locations, record lengths and number of stations used.

https://doi.org/10.5194/essd-2021-339
Preprint. Discussion started: 29 November 2021



615 **Figure 2: Input data for daily EMO-5 vapour pressure, solar radiation and wind speed grids in terms of station locations, record lengths and number of stations used.**

https://doi.org/10.5194/essd-2021-339
Preprint. Discussion started: 29 November 2021

Open Access Earth System Science Data Discussions

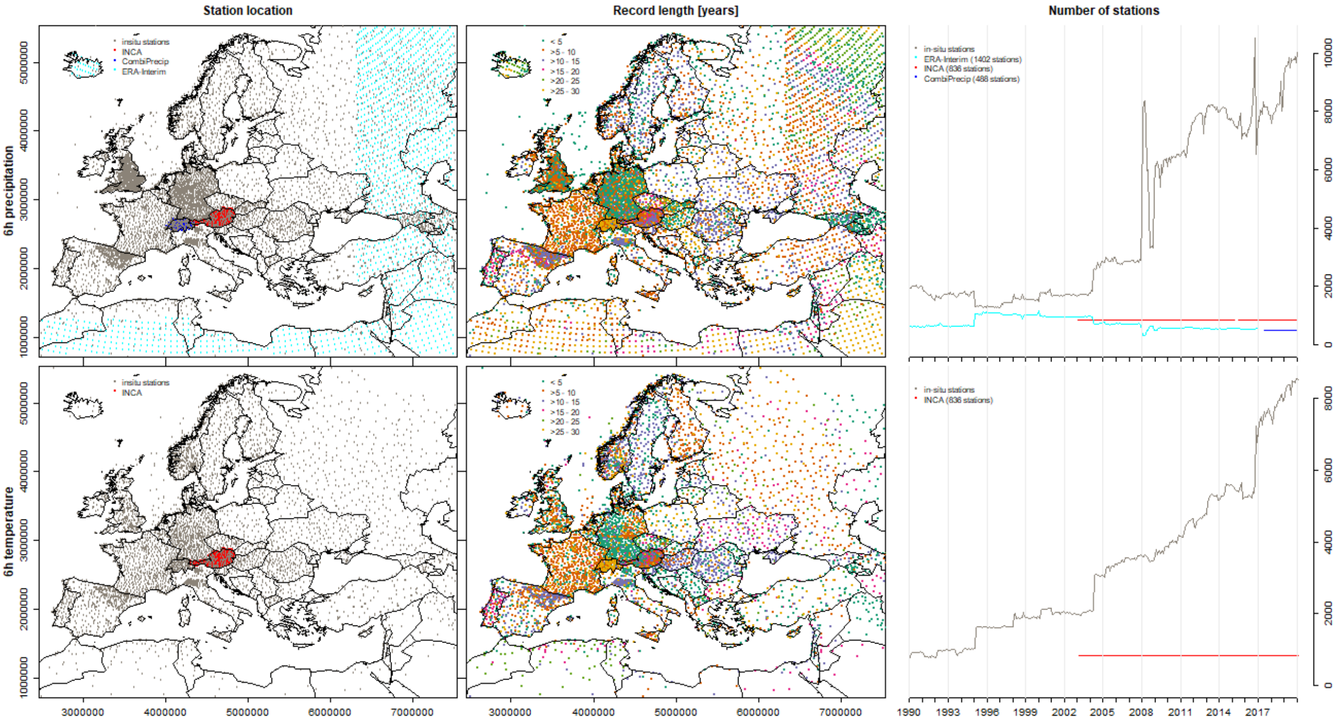


Figure 3: Input data for 6-hourly EMO-5 precipitation and temperature grids in terms of station locations, record lengths and number of stations used.

https://doi.org/10.5194/essd-2021-339
Preprint. Discussion started: 29 November 2021

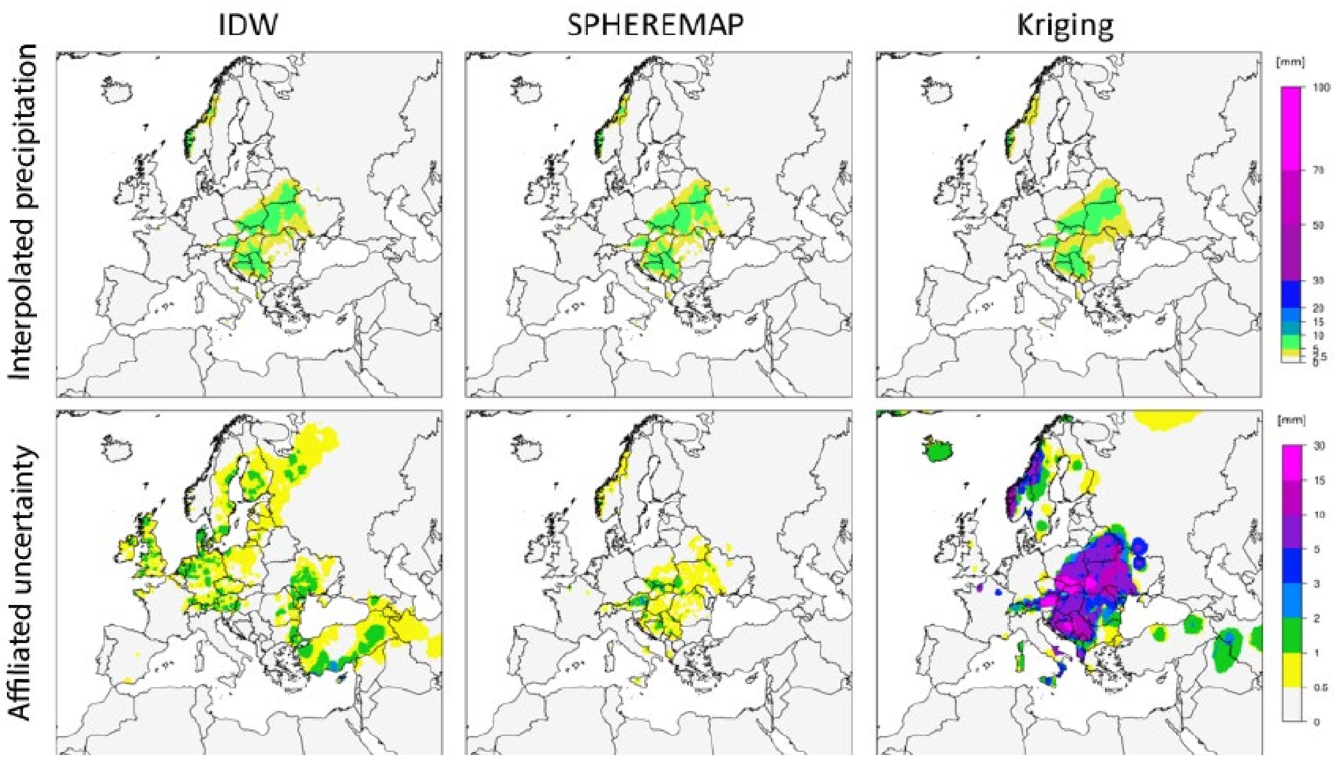


Figure 4: Gridded daily precipitation data (top row) and the corresponding uncertainty fields (bottom row) for the 15.05.2014 as an example for IDW (left), SPHEREMAP (middle) and Kriging (right).

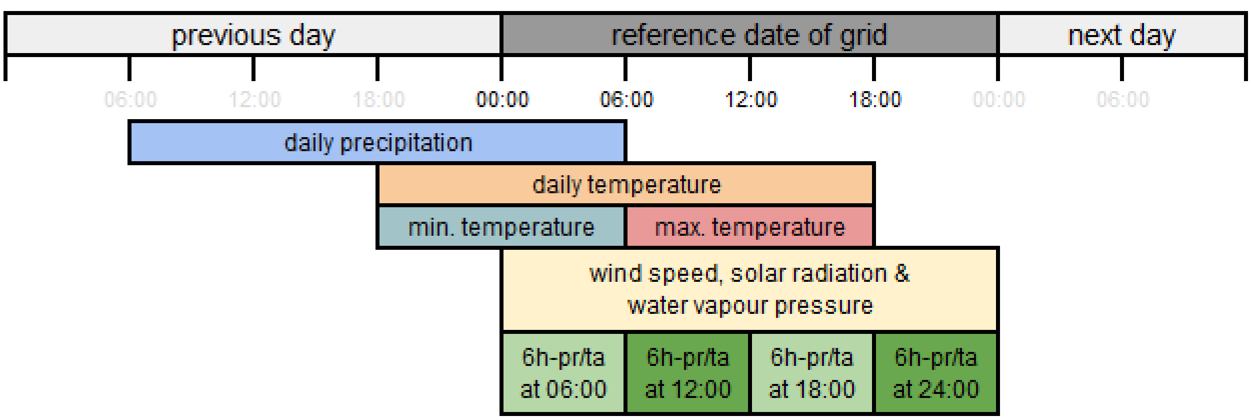


Figure 5: Aggregation i.e. reference period of the EMO-5 grids per variable.

https://doi.org/10.5194/essd-2021-339
Preprint. Discussion started: 29 November 2021

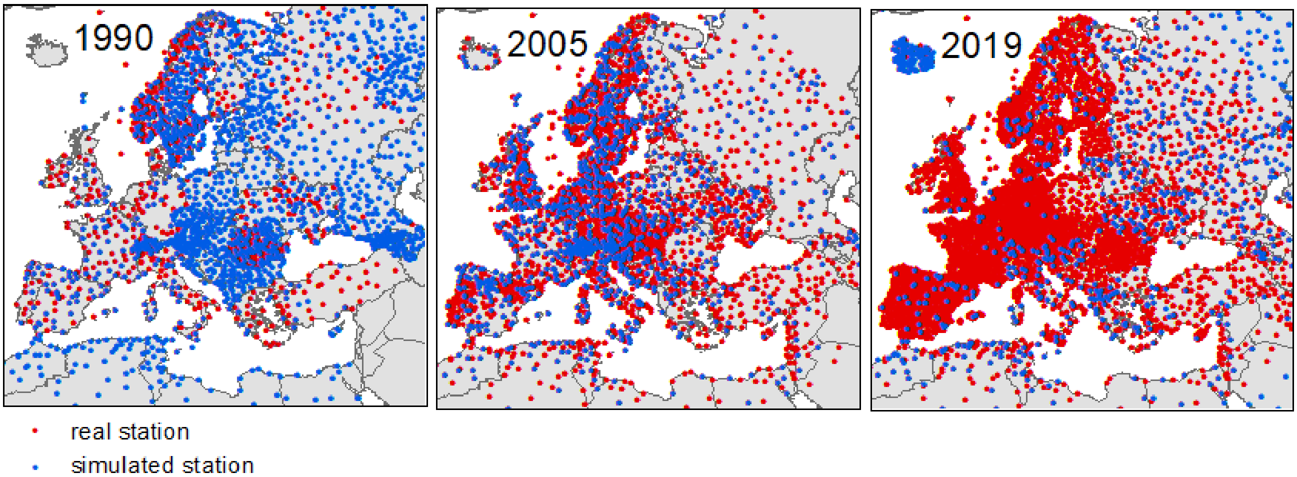


Figure 6: Coverage of existing and simulated stations used for the 6-hourly temperature fields for 1990, 2005 and 2019.

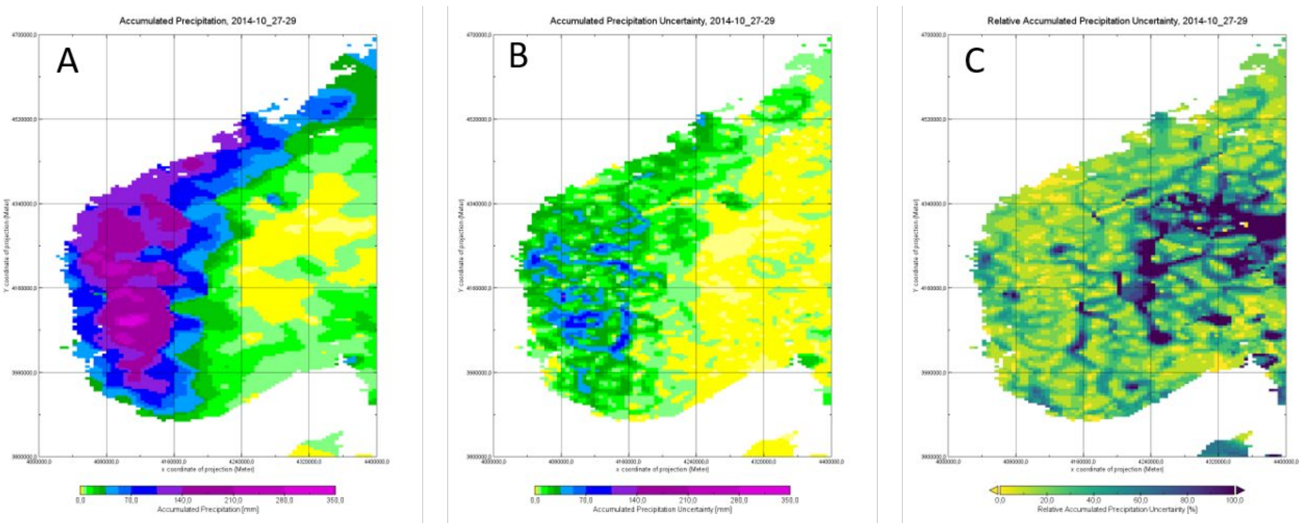


Figure 7: Example of interpolated precipitation totals and associated uncertainty for extreme event No. 3 in Section 5.3. (A) Accumulates precipitation totals from 2014/10/27 to 2014/10/29 (three days), (B) accumulated uncertainty of the interpolated precipitation totals and (C) the percentage of the accumulated uncertainty in relation to the accumulated precipitation totals.

https://doi.org/10.5194/essd-2021-339
Preprint. Discussion started: 29 November 2021

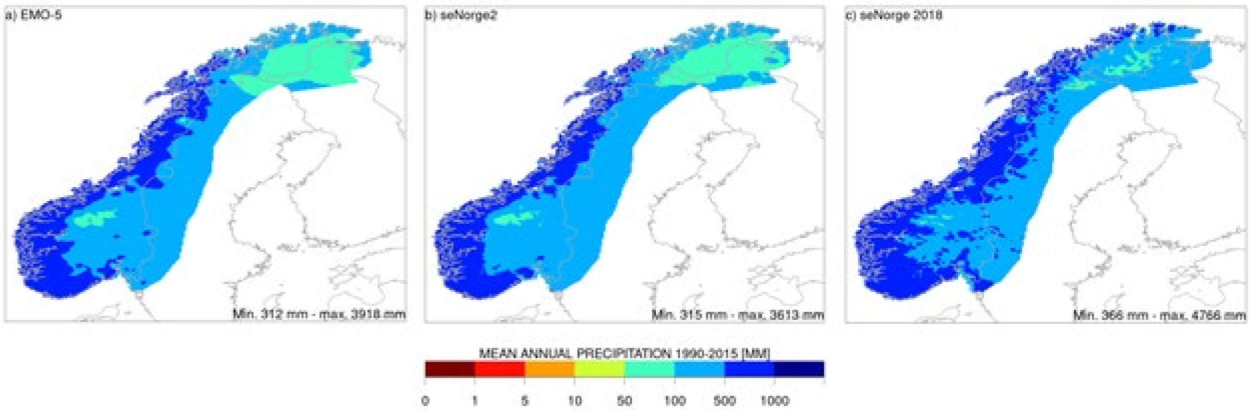


Figure 8: Mean annual precipitation [mm] for the (a) EMO-5, (b) seNorge2, and (c) seNorge 2018 data sets for the joint period 1990-2015. seNorge data are bi-linearly interpolated to the coarser EMO-5 resolution. EMO-5 was cropped to the same spatial extent as the seNorge data sets.

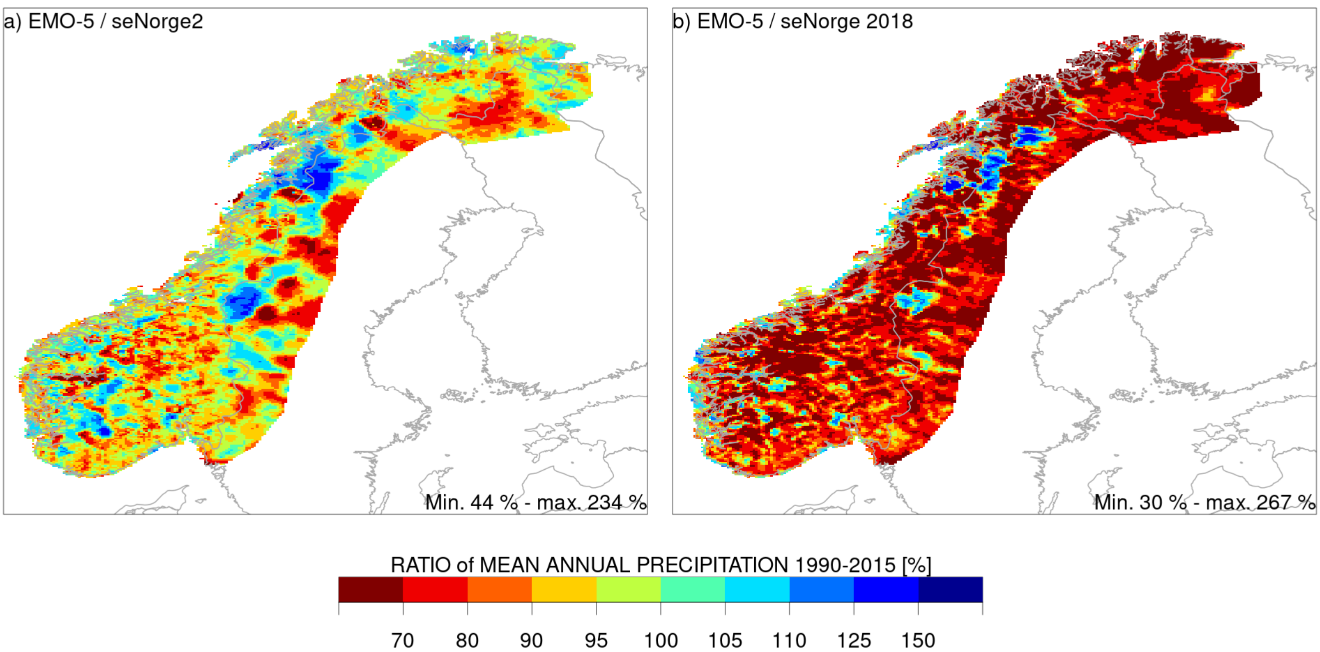


Figure 9: Based on Figure 8 data. Ratio of seNorge2 (a) and seNorge 2018 (b) to EMO- 5 [%].

https://doi.org/10.5194/essd-2021-339
Preprint. Discussion started: 29 November 2021

Open Access
Earth System Science Data
Discussions

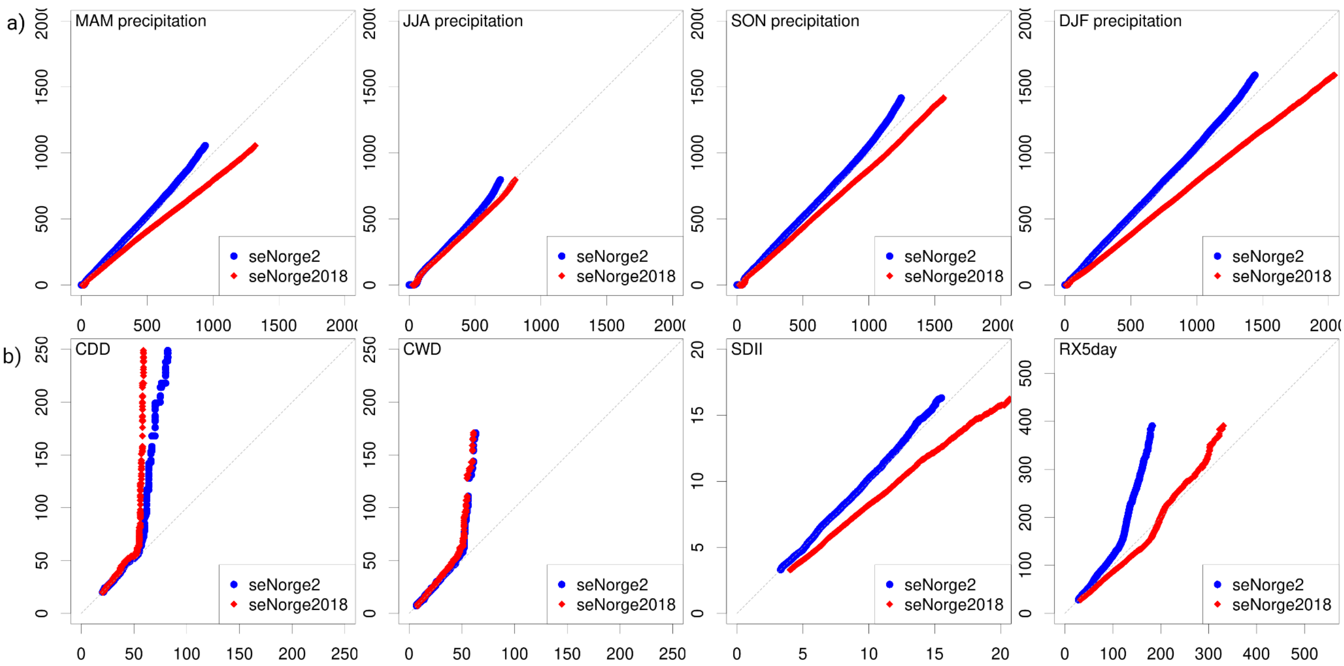
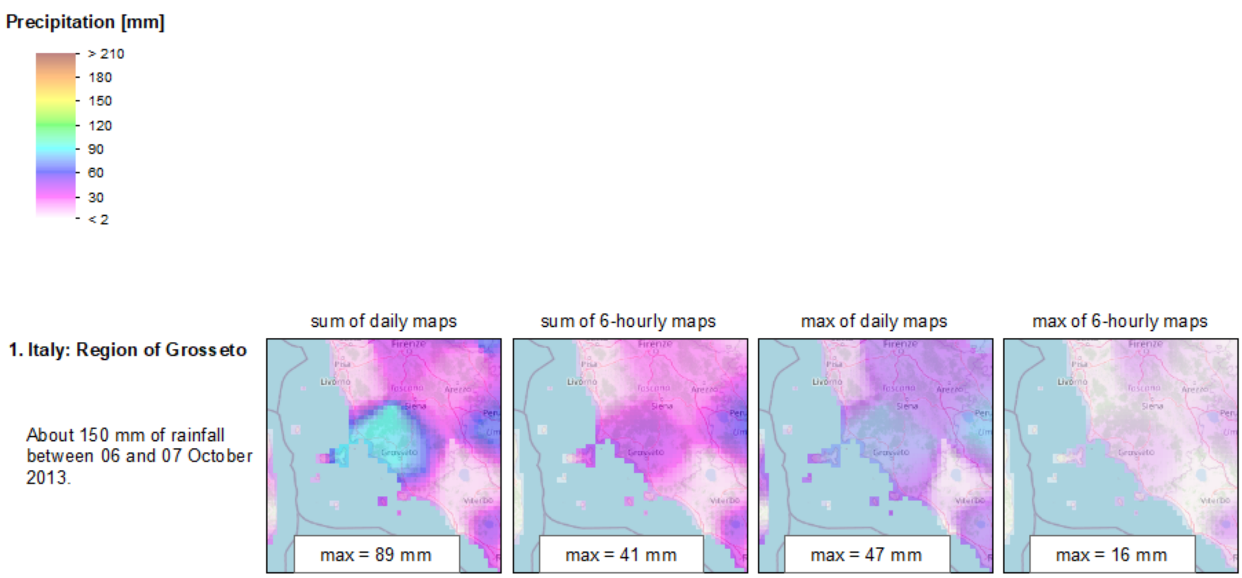
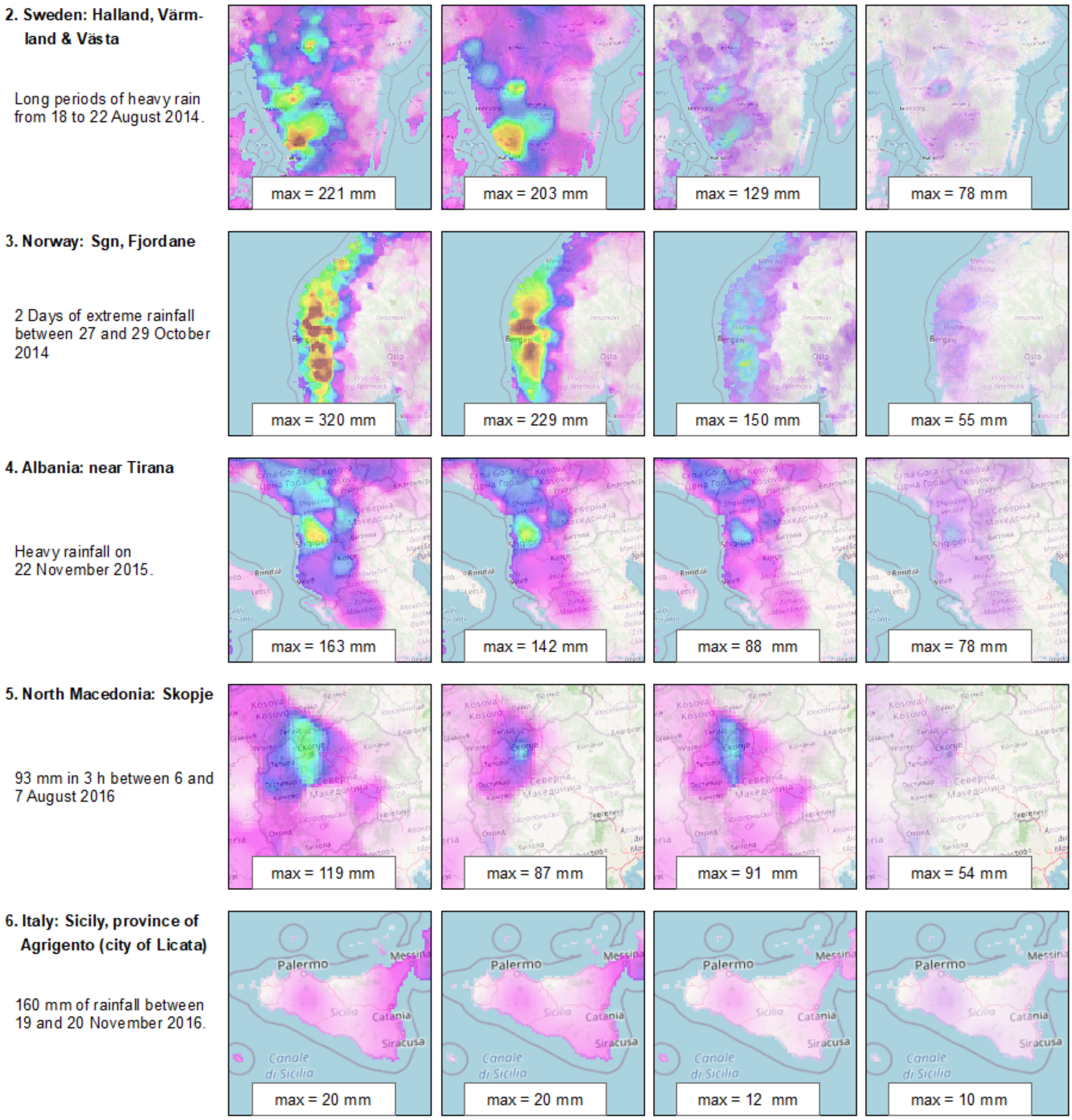


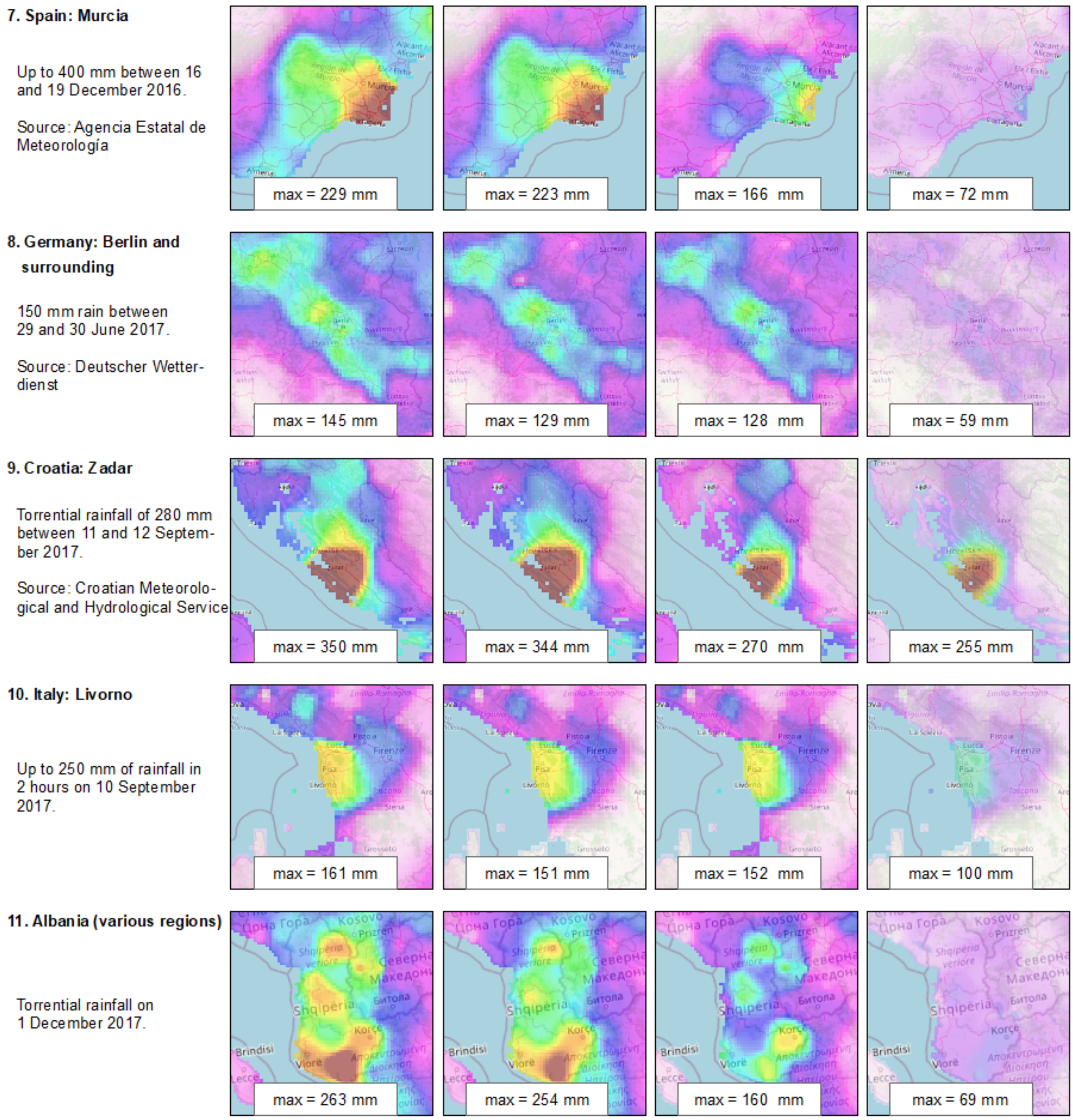
Figure 10: Quantile plot of a) seasonal precipitation, where each point represents the sum of a grid value over 3 months and b) extreme value indices (ETCCDI) values, but these correspond to the maximum of the entire time series. Each plot shows seNorge2 (●) and seNorge2018 (♦) values on the x-axis and EMO-5 values on the y-axis in mm. For a better visual overview, the highest 0.01% of the data have not been shown.



https://doi.org/10.5194/essd-2021-339
Preprint. Discussion started: 29 November 2021

Open Access Earth System Science Data Discussions




7. Spain: Murcia
Up to 400 mm between 16 and 19 December 2016.
Source: Agencia Estatal de Meteorología
max = 229 mm
max = 223 mm
max = 166 mm
max = 72 mm
8. Germany: Berlin and surrounding
150 mm rain between 29 and 30 June 2017.
Source: Deutscher Wetter-dienst
max = 145 mm
max = 129 mm
max = 128 mm
max = 59 mm
9. Croatia: Zadar
Torrential rainfall of 280 mm between 11 and 12 September 2017.
Source: Croatian Meteorological and Hydrological Service
max = 350 mm
max = 344 mm
max = 270 mm
max = 255 mm
10. Italy: Livorno
Up to 250 mm of rainfall in 2 hours on 10 September 2017.
max = 161 mm
max = 151 mm
max = 152 mm
max = 100 mm
11. Albania (various regions)
Torrential rainfall on 1 December 2017.
max = 263 mm
max = 254 mm
max = 160 mm
max = 69 mm

https://doi.org/10.5194/essd-2021-339
Preprint. Discussion started: 29 November 2021

Open Access Earth System Science Data Discussions

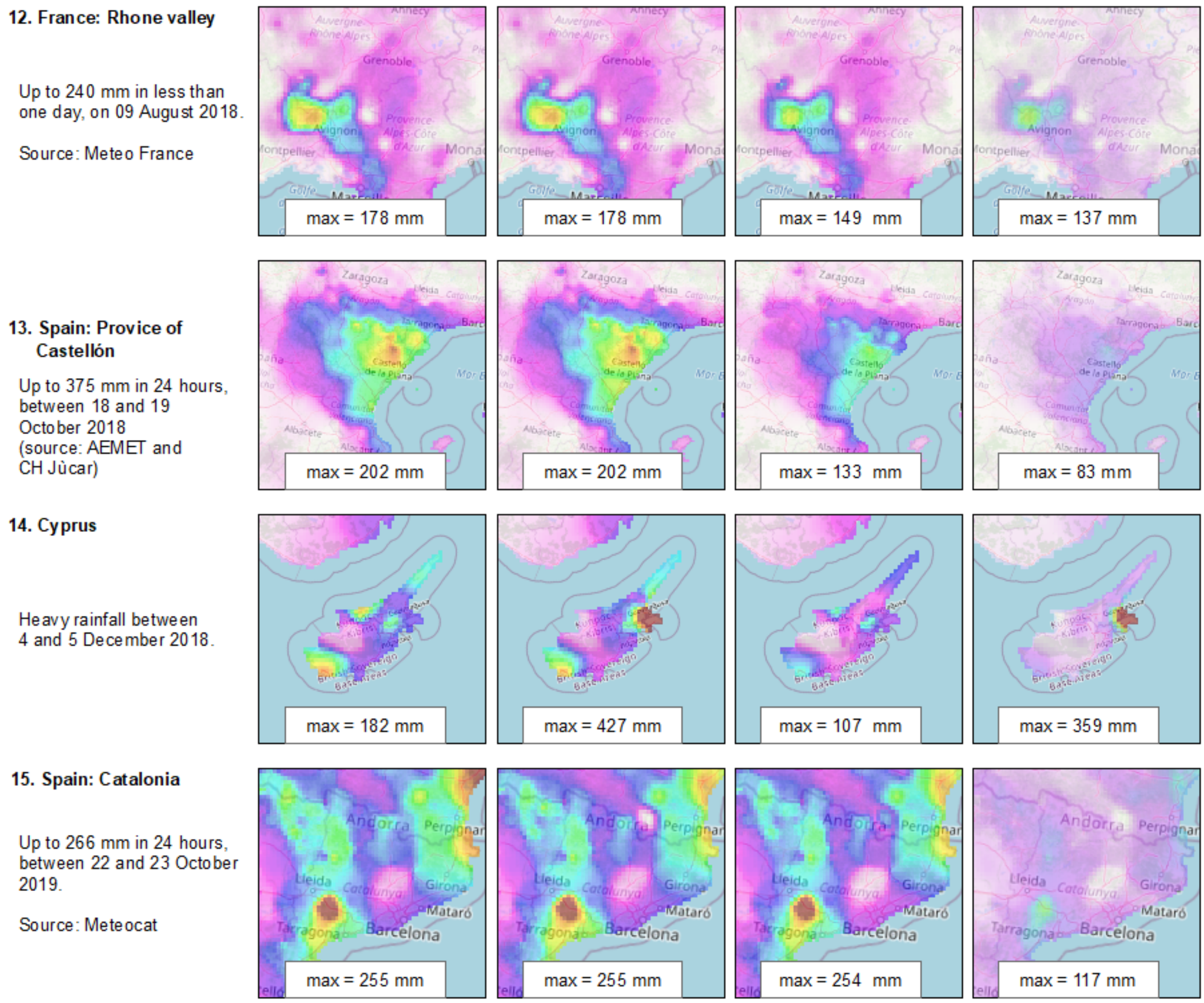


Figure 11: Semi-qualitative evaluation of 15 high intensity precipitation events through comparing information reported on FloodList (on left) with the footage of the 6-hourly and daily precipitation grids (sums and maximum) (base map from © OpenStreetMap contributors 2020. Distributed under the Open Data Commons Open Database License (ODbL) v1.0)

https://doi.org/10.5194/essd-2021-339
Preprint. Discussion started: 29 November 2021

Tables

Table 1: Overview of the five gridded meteorological data sets that form the input for EMO-5.

Gridded data set (reference)	Description
ZAMG-INCA (Haiden et al., 2011)	A near real-time high-resolution (1x1 km) multi-variable meteorological data set covering the whole of Austria. The MDCC imports from the INCA data set 6-hourly precipitation and 6-hourly mean temperature information at 836 pre-selected locations into the MDCC data collection. The data collection is done in near real-time since 2009, whereas historical data from 01.01.2003 onwards have been loaded as a bulk order.
COMBI-PRECIP (Sideris et al., 2014)	A near real-time radar / rain-gauge product for hourly precipitation covering the whole of Switzerland as well as bordering regions. 1-hourly precipitation data is imported into the MDCC data collection in near real-time at 488 virtual stations across Switzerland since 01.06.2017.
EURO4M-APGD (Isotta et al., 2014)	A daily precipitation data set from 1971 till 2008 covering the Alps and adjacent flatlands (area between 4.8 and 17.5°E as well as 43 to 49°N). The daily precipitation is imported into the MDCC data collection at 4960 virtual stations from 01.01.1971 till 31.12.2002. From 01.01.2003 the number of virtual stations imported to the MDCC data collection dropped to 4123 stations, as 837 stations over Austria were excluded due to the availability of the higher resolution INCA data covering that area.
CarpatClim (Antolović et al., 2013; Spinoni et al., 2015)	Another historical daily precipitation product, covering the area between 44 and 50°E as well as 17 to 27°N (Hungary, Serbia, Romania, Ukraine, Slovakia, Poland, Czech Republic, Croatia) at a horizontal resolution of 0.1°. The data is imported into the data collection for the entire historical period from 1970 till 2010 as 2946 virtual stations.
ERA-Interim/Land (Dee, 2011; Dee et al., 2011)	A global multi-variable reanalysis data set for all land surface areas from 1979 till October 2016 at a 0.75°x0.75° spatial resolution. The import of ERA-Interim/Land data has been limited to peripheral areas for which the coverage with in situ stations in the MDCC database is very low, such as Iceland, North Africa and the eastern part of the EFAS domain (Near East, Caucasus, Russia). For those areas, 6-hourly precipitation was imported for 1402 virtual stations.

Table 2: Specifications for the validation rules as applied on the EMO-5 source data

Variable	Monthly validation (Rule 2)	Cross validation (Rule 3)	Min/max validation (Rule 4)		Rate of change validation (Rule 5)
			Minimum threshold	Maximum threshold	
precipitation [mm]	yes		0	• 125 (15 min)* • 200 (30 min) • 250 (1 h) • 350 (3 h) • 425 (6 h) • 475 (9 h) • 500 (11 h) • 525 (15 h) • 550 (18 h) • 600 (24 h)	
Mean temperature [°C]	-		-45	+43	• 10 (15, 30, 60 min)** • 20 (3, 6 h) • 25 (12 h) • 30 (24 h)
Min temperature [°C]	-		-50	+35	
Max temperature [°C]	-		-40	+50	
Vapor pressure [hPa]	yes		-	-	
Solar radiation [W/m²]	yes		0	1360	
Wind speed [m/s]	yes	Wind speed (WS) depending on Wind direction (WD) • IF WS = 0 AND WD ≠ 0 THEN WS suspect • IF WS ≠ 0 AND WD = 0 THEN WS suspect	0	45	

* the maximum threshold for precipitation depends on the aggregation interval

** the maximum change of temperature [K] depending on the observational interval (given in parenthesis)

https://doi.org/10.5194/essd-2021-339
Preprint. Discussion started: 29 November 2021

Open Access Earth System Science Data Discussions

Table 3: Overview of the three spatial interpolation schemes that were evaluated for the purposes of EMO-5.

#	Interpolation scheme	Description
1.	**Inverse Distance Weighting (IDW)**	IDW is a simple and robust scheme with low computational cost. It is a purely geometric scheme based on the assumption that the closer the meteorological station is to the grid cell centre, the more related it is to its actual value. Mathematically, this is expressed by assigning weights to the surrounding stations proportional to their distance d, e.g. $1/d^2$.
2.	**Modified SPHEREMAP**	The original SPHEREMAP (Willmott et al., 1985) is the adaptation to spherical coordinates of Shepard's inverse distance weighting (Shepard, 1968), which is an extension to the IDW scheme described above. Interpolation weights decrease with increasing distance between grid centre and meteorological station, as in IDW, but the equation for the calculation of the interpolation weight depends also on the distance and number of available input stations. Furthermore, the interpolation takes the clustering of stations into account, so the weights of clustered stations were reduced in order not to overweight these data. As the original SPHEREMAP scheme would lead to a neglect of many stations in regions with a high station density we adapted the algorithm. Previously, if at least one station was found within the smallest search radius "epsilon", then this station or the mean of the stations within "epsilon" was utilised and the station outside the "epsilon" neglected. Now, the "epsilon" is set as 1/20 of the initial search radius and the distance from stations within the radius is set to "epsilon" to avoid an overweighting of the nearest station(s). With these modifications the utilisation of at least four stations per grid point is assured. The maximum number of stations used for interpolation is set to 10.
3.	**Ordinary Kriging**	Ordinary Kriging (Krige, 1966) is an advanced geostatistical method based on correlations between observations. The interpolation weights, which are created by means of the variograms, make use of observation data. Briefly, variograms sort the variance between observations by the distance between these observations were taken. Several approaches can be used to compute these variograms, such as calculations of variograms for each station and time separately, or climate zone dependent variograms, but here utilizing one global variogram for all interpolations, as is utilized at the Global Precipitation Climatology Centre / GPCC (Schamm et al., 2014).

695

https://doi.org/10.5194/essd-2021-339
Preprint. Discussion started: 29 November 2021

Table 4: Summary of the error measures for the three interpolation schemes. (Best values are in bold)

	ME			**MAE**			**MSE**		
	IDW	**SP**	**KRI**	**IDW**	**SP**	**KRI**	**IDW**	**SP**	**KRI**
Precipitation	0.89017	**-0.01585**	-0.02312	2.3233	**1.35704**	1.4019	92192.22	12.88	**12.304**
Min temp	0.06418	**0.04342**	0.060102	**1.6015**	1.6164	1.6225	**5.2379**	5.592	5.3695
Max temp	0.035612	0.045672	**0.00394**	**1.7648**	1.7781	1.7878	**7.2925**	8.0828	7.5285
Wind speed	0.026899	**0.007977**	0.0322	**0.9628**	1.02	0.9706	**1.9856**	2.31	2.0056
Water vapour pressure	0.010876	**0.002848**	0.017395	0.8525	**0.8415**	0.85295	1.9404	1.9568	**1.9287**
Radiation	22.1408	17.2062	**15.3328**	**2151.65**	2283.68	2152.9	**8686705**	10074971	8690056

Table 5: Overview of EMO-5, seNorge2 and seNorge2018 data sets based on the common period 1990-2015 [mm]. The order of calculation is first to calculate the parameters (mean, maximum) for the grids, followed by the mean over the resulting grid.

	EMO 5					**SeNorge2**					**SeNorge 2018**				
	Ann	MAM	JJA	SON	DJF	**Ann**	MAM	JJA	SON	DJF	**Ann**	MAM	JJA	SON	DJF
mean	968	183	248	279	255	948	179	247	274	247	1163	227	266	328	340
max	1371	346	423	496	467	1291	317	406	471	433	1601	424	446	573	603

https://doi.org/10.5194/essd-2021-339
Preprint. Discussion started: 29 November 2021

Open Access Earth System Science Data Discussions

Appendix 1

Table A1: EMO-5 data provider (in grey shaded the providers of gridded data).

Data provider acronym	Data provider name (note: name of data sets shaded in grey)	number of stations within the grid domain	precipitation	temperature	wind speed	solar radiation	vapor pressure
MARS	Wageningen Environmental Research (ALTERRA)	5442	x	x	x	x	x
EURO4M-APGD	European Reanalysis and Observations for Monitoring (EURO4M-APGD)	4960	x	-	-	-	-
MeteoConsult	MeteoConsult	4400	x	-	-	-	-
DWDSynop	Deutscher Wetterdienst	3446	x	x	x	-	-
CarpatClim	CarpatClim - Climate of the Carpathian region	2946	x	-	-	-	-
ERA Interim	European Centre for Medium-range Weather Forecasts (ECMWF)	1402	x	-	-	-	-
NMI	Norwegian Meteorological Institute	1237	x	x	x	-	-
ECA	European Climate Assessment and data set (ECA)	1201	x	-	-	-	-
EA	UK Environment Agency (EA)	897	x	-	-	-	-
AMDASynop	Deutscher Wetterdienst	849	x	x	x	-	-
ZAMG	Zentralanstalt für Meteorologie und Geodynamik	836	x	x	-	-	-
CombiPrecip		488	x	-	-	-	-
SAIH Ebro	Confederacion hydrografica del Ebro	386	x	x	?	x	-
ARPASIM	Agenzia Regionale per la Prevenzione e l'Ambiente dell'Emilia-Romagna	145	x	x	-	-	-
MeteoSwiss	MeteoSchweiz	140	x	x	x	x	-
FMI	Finnish Meteorological Institute (FMI)	110	x	x	-	-	-

https://doi.org/10.5194/essd-2021-339
Preprint. Discussion started: 29 November 2021

Open Access Earth System Science Data Discussions

HMS	Hungarian Meteorological Service	97	x	x	-	-	-
IPMA	Institute for Ocean and Atmosphere, Portugal	96	x	x	x	x	-
SHMU	Slovak Hydro-Meteorological Institute	72	x	x	-	-	-
CHMI	Czech Hydro-Meteorological Institute	69	x	x	-	-	-
IMGW	Institute of Meteorology and Water Management - National Research Institute (Poland)	50	x	x	-	-	-
METIE	Met Éireann	32	x	x	x	-	-
ARSO	Slovenian Environment Agency	21	x	x	x	x	-
KHMI	Hydrometeorology Institute of Kosovo	4	x	x	-	-	-