Dear Reviewers,

Once again, thanks for your comments and suggestions. You may have noticed that we have answered to your comments in the interactive discussion and for your convenience we copy here the link:
https://www.earth-syst-sci-data-discuss.net/essd-2019-43/

We have modified the structure of the paper, according to your suggestions. The revised manuscript is organized as follows:
- Introduction: in order to describe the original aspects of our research, we need to introduce concepts such as the effective resolution of a dataset and the definitions of spatial scales (in a way that is useful for our purposes). Those paragraphs may also be regarded as methods, nonetheless we believe they belong in the introduction.
- Data: this Section has been split into three sub-sections. First, the observations are described, together with the procedure used to post-process precipitation measured data. Second, we describe the reference fields for the spatial interpolation of precipitation. Third, IDI is introduced and discussed here, because we intend
- Methods: the core of our paper in the spatial analysis. For this reason, we believe the methods should be focused on spatial interpolation (e.g., post-processing of precipitation measurements is part of the Data section).
- Example application for precipitation: new section. The spatial analysis of precipitation is here described step-by-step. As stated in the last paragraph of the Introduction, examples of temperature analyses for individual cases can be found in a previous paper. We have put a lot of effort in the characterization of the analysis products in the verification section.
- Verification: the name of this section has been changed from Results to Verification. The text has been revised and the section has been split into Verification and discussion. We have added the paragraph on Precipitation and effective resolution.
- Discussion: new section. Some important issues have been discussed in detail, such as the TG, TX and TN cross-checking.
- Conclusions.

Additional calculations have been performed, to either support our conclusions or point out new features of seNorge_2018. In particular: the effective resolution of RR has been assessed; the significance of land-area-fraction in reducing the uncertainties in the analysis of coastal stations has been verified through cross-validation.

Several figures have been added and most of the original figures have been modified.

Point-by-point response to your reviews follows. The Reviewers' comments are reported in Italic.

Best Regards,
Cristian Lussana on behalf of the Authors

Authors' Response to Reviewer 1

Main concerns

*) The story line is sometimes hard to follow and the readability would improve with some more illustrations. For instance: a brief introduction to optimal interpolation could be helpful on page 5. Another example are the scales you introduce on page 9 (line25-27). A visualisation of some of these scales would help in guiding the reader to understand the approach. Similarly on page 10, line 15-16: can you visualize this scale somehow and show how this scale changes over time?

Reply: We have completely revised the structure of the manuscript. We have included an example of application for RR and the figure mentioned in your comment has been placed there.

Figure 1: I understand that the explanation of the colour coding in fig 1a and 1b is complex, but now the reader has to read a substantial part of the article first before he/she understands what you are plotting here. An intuitive explanation for IDI might help for the reader who has a look at the figures first before deciding to read the paper. Explanation of abbreviations IDI and CV-IDI also helps. In the precipitation plot I'm missing the station locations.

Reply: Fig. 1 has been modified such that now the issues raised by the Reviewer should be solved.

*) A smaller issue is the structure of the text, a critical look would help here. For Instance, on page 3, line 10 you start to claim that your approach will capture field variability at unresolved spatial scales. The next line is not an explanation of how this is achieved, but deals with something quite different. It is until line 17 that the reader is informed how you take-in the information on the unresolved spatial scale. Another example is on page 5, line 30. You write '...and Fig 1 shows those regions'. It would help if you guide the reader more explicitly where to look in Fig 1 (which regions/colours,which subfigure).
Reply: we have revised the manuscript as suggested.

*) Relating to the interpretation of the results: Figure 2 shows that the analysis of TN is increasing with increasing CV-IDI Both the background and CV-analysis are decreasing with CV-IDI, but for summer this is not seen in the analysis. In winter this effect is absent as well. My first guess was that this might be the influence of the urban stations, picking up the urban heat island effect. Can you comment on this?

Reply: with reference to the interactive comments, we have included a comment on this topic in the manuscript (Section "Discussion").

*) In addition: In the introduction there is a paragraph concerning the effective resolution of grids. This makes me curious about the effective grid resolution of seNorge_2018 compared to seNorge2. Is there a way to quantify this? This is an interesting aspect, since the number of station observations (~density of network) is similar in both datasets.

Reply: We have used a scale-decomposition approach (based on 2D wavelet) to the daily precipitation fields of seNorge_2018 and seNorge2. The results have been included in the Section "Discussion", together with a dedicated Figure.

Other points the authors may want to look into:
*) On page 4 (bottom) you describe the increase in station density and that many of these stations are installed in cities and villages. I was wondering if this aspect would give you a possibility to assess the Urban Heat island effect in Norway's larger cities? A comment on this would be interesting.

Reply: see the interactive comments.

*) Page 5, line 25. Wouldn't the complexity of the topography be a relevant function here, and if so, have you looked into topography complexity? (slope, aspect, elevation)

Reply: see the interactive comments.

*) Page 7, line 4-5. It is a good thing that physical consistency is enforced. A brief explanation of how this is done is helpful, i.e. are you simply setting tn to tg where you find that tn is larger than tg, or is there a slightly more sophisticated approach?

Reply: We have added some text to the manuscript to describe the check.

*) On page 8, line 4: what is the motivation to choose a 50x50 grid? Where there any sensitivity analysis to support the choice?

Reply: We have added some text in the Section "Methods" of the manuscript to better describe why we have chosen a 50x50 grid.

*) The usage of 100 scales with a minimum of 2 and a maximum of 1400 km is unclear to the reviewers. Can you visualize some of these scales? Is there a more graphical way to explain how you use these successive scales for downscaling.

Related to this, can you visualize the critical scale mentioned on page 10, line 15? Such a graphical representation convinces readers about the innovation of your method.

Reply: A graphical representation has been added.

*) page 11, line 24: can you comment if you think there are other ways that might alleviate this problem with TN without having to install new stations?

Reply: see the interactive comments. We have added a discussion on this point in the text.

*) page 12, line 5: here you claim that the addition of the land area fraction in equation 7 improves the temperature fields. What you are showing is the difference. Intuitively I see where you are going, but showing an improvement requires the cross validation, and the reviewer has not seen evidence that the new dataset improves considerably along the coastline.

Reply: We have included in the text the verification based on cross-validation and considering coastal stations only.

*) page 12, line 31: I agree with your statement, but the reverse does not seem to be true. In the Oslo fjord the station density is very high but the TN quality is as low as in less dense regions.

Reply: see the interactive comments.

*) The results section starts with the description of the CV. There are some concerns about the validation methods. A LOOCV or random sampling with k=folds does assume data points are spatially independent. This does not hold for data dense regions, moreover these data points will be predicted accurately due to their spatial dependence (especially using LOOCV). It is expected that the current approach will result in an underestimation of prediction errors. A way of making the validation procedure less spatially dependent (and less computationally expensive) could be to split the data point into k-equal area folds.

Reply: see the interactive comments. We have modified the text so to take into account the Reviewer's remark.

*) page 6, line 5,8: What is the motivation to choose the numerical boundaries for CV-IDI to indicate data dense and sparse regions?

Reply: we have included an explanation in the manuscript.

) page 10, line 19: This is not a general description of CV but LOOCV.

Reply: The text has been modified.

) General remark on figures: in most of the figures I'm missing the subfigure annotations (a,b,c). Please include a raster grid with lat/lon for the spatial plots. For regularly gridded raster plots one scale bar is sufficient.

Reply:

) Figure 4: Good to include lat/lon averaged differences, am I assuming correctly that the grey area are the min/max values (between -4 and +1.5)? The lateral and bottom panels y-axis temperatures are hard to read. This does not support the text on page11, line 28, which suggests that almost all differences are between -2 and +1.

Reply: We have modified the text accordingly.

) page 7, line 18: what does the i mean in the definition of G and S as these latter quantities appear not to be related to grid point i?

Reply: G and S are related to the i-th gridpoint because for each gridpoint a different horizontal decorrelation length is used.

) Table 1: The caption states that the third station is used while in the text (page 8, line33) the average distance to the nearest four stations is used.

Reply: we have modified the text so to correct for the mistake.

Very minor issues

Reply: thanks for pointing out the following issues.

) page 3, line 27: "Finally, Section 4..."→This suggests there is no chapter 5.

) page 7, line 18: typo in gridpoint

) page 10, line 7: I assume "to have the same error"

) page 13, line 8: Shouldn't 0.4 be 40%?

*) page 13, line 18: I guess this should be "paper of Lussana"

*) page 17, missing pages in the Reistad citation

*) Figure 7: Using black in the color scale is inconvenient since country borders and coastlines are also black.

Authors' Response to Reviewer 2

The authors present a new version of Norwegian national daily temperature and precipitation interpolated daily fields from the latter half of the 20thCentury to date. The data product shall, undoubtedly, constitute a valuable national resource for decision makers within Norway and the broader Nordic region. To build confidence in the product peer review is certainly a necessary condition and thus I see eventual publication as important. In reviewing the discussion paper there are a number of issues thatI believe the authors must address prior to acceptance.

Firstly, the paper structure requires significant work. The introduction mixes methods and discussion. The data section has a significant amount of methods in it and does not clearly denote the various observational / model sources used. Presently there is very limited description / analysis of the derived spatial fields and characterisation which would be important for users. Finally, there is no discussion section. My suggestion would be to substantively restructure the paper for readability into sections that go:
•introduction,
•data,
•methods,
•product analysis(including showing some example applications),
•verification,
•discussion and
•conclusion.
Then significant effort is required to shuffle content around to fit that structure, ensuring that relevant text ends up in the appropriate section.Given the need to restructure the paper I shall not point out minor typographical issues in the expectation that they may not persist under any revised structure.A number of other minor points should also, naturally, be resolved by undertaking such a restructure so I do not make these further here.Such a restructuring to my view is essential prior to acceptance.

Reply: we have modified the structure of the paper.

In terms of the methods there is a significant issue in offsetting Tx/Tn from Tg by 12 hours. If Tg is the mean of 06 to 06 but Tx and Tn are maximum and minimum between 18 and 18 it is physically impossible to robustly assess consistency. This has been shown in e.g. GHCND and can follow from several toy examples you may wish to play with whereby for example a very strong warm front passes through at midnight which would be seen in one day for Tx and Tn but another day for Tg and may lead to an over-propensity of flagging good data as dubious accordingly.This

propensity will vary seasonally (higher in winter half year) and geographically (higher further north/ inland) where both diurnal structure decreases and synoptic variability increases.Significant justification would be required for maintaining the use of days offset by 12 hours for the three temperature elements and my strong recommendation would be to align these to the same time which would greatly simplify the analysis and assure better geophysical consistency with fewer false flags.It would also aid usability considerably to align the times for all 4 elements. So, whether you choose 06 to 06, 18 to 18 or some other times I would very strongly urge aligning the times used to define the day here which would enable greater usability and improved cross-checking.

Reply: We have to use two different definitions of day, as reported in the interactive discussion. This fact is stated explicitly in the definition of the variables (beginning of Section 2.1).

The authors make a throw away remark at the end of page 2 regarding suitability for long-term trend characterisation which seems to rely upon findings of a prior analysis. It is unclear whether the findings would persist into the present dataset in the Norwegian context. It is necessary, in my view, to show this and the suggested change in overall paper structure should facilitate this.

Reply: we have rephrased the statement so to avoid confusion.

In the methods Xi is used twice, one should be Xj. Then the same overhat nomenclature is used to denote both a point estimate and a spatial scale. This is very confusing to the reader in what is already a very statistically dense paper. Assuming that the average ESSD paper is not a statistician it would be very useful to simplify where possible the discussion of methods and certainly to use unique notations when talking about distinct things so as to not confuse unnecessarily your readers. Overall, a reduction in the number of equations would likely serve the ESSD readership.

Reply: We have revised a bit the notation. We believe it is not possible to reduce the number of equations without compromising the possibility to reproduce our methods.

The right hand panel of figure 6 uses a non-intuitive colour scale whereby wetter values are red and drier values blue(as I understand this panel at least). If I am correct it would be advisable to flip the colour bar so that the colours intuitively map

to wet / dry rather than doing so counter-intuitively. If I am wrong then an improved explanation is required.

Reply: The figure has been modified as suggested.

seNorge_2018, daily precipitation and temperature datasets over Norway

Cristian Lussana[1], Ole Einar Tveito[1], Andreas Dobler[1], and Ketil Tunheim[1]

[1]Norwegian Meteorological Institute, Oslo, Norway

Correspondence: Cristian Lussana (critianl@met.no)

Abstract. seNorge_2018 is a collection of observational gridded datasets over Norway for: daily total precipitation; daily mean, maximum and minimum temperatures. The time period covers 1957 to 2017, and the data are presented over a high-resolution terrain-following grid with 1 km spacing in both meridional and zonal directions. The seNorge family of observational gridded datasets developed at the Norwegian Meteorological Institute (MET Norway) has a twenty-year long history and seNorge_2018 is its newest member, the first providing daily minimum and maximum temperatures. seNorge datasets are used for a wide range of applications in climatology, hydrology and meteorology. The observational dataset is based on MET Norway's climate data, which has been integrated by the "European Climate Assessment and Dataset" database. Two distinct statistical interpolation methods have been developed, one for temperature and the other for precipitation. They are both based on a spatial scale-separation approach where, at first, the analysis (i.e., predictions) at larger spatial scales are estimated. Subsequently they are used to infer the small-scale details down to a spatial scale comparable to the local observation density. Mean, maximum and minimum temperatures are interpolated separately, then physical consistency among them is enforced. For precipitation, in addition to observational data, the spatial interpolation makes use of information provided by a climate model. The analysis evaluation is based on cross-validation statistics and comparison with a previous seNorge version. The analysis quality is presented as a function of the local station density. We show that the occurrence of large errors in the analyses decays at an exponential rate with the increase in the station density. Temperature analyses over most of the domain are generally not affected by significant biases. However, during wintertime in data-sparse regions the analyzed minimum temperatures do have a bias between $2°C$ and $3°C$. Minimum temperatures are more challenging to represent and large errors are more frequent than for maximum and mean temperatures. The precipitation analysis quality depends crucially on station density: the frequency of occurrence of large errors for intense precipitation is less than 5% in data-dense regions, while it is approximately 30% in data-sparse regions. The open-access datasets are available for public download at: daily total precipitation (DOI:https://doi.org/10.5281/zenodo.2082320 **?**) ; daily mean (DOI:https://doi.org/10.5281/zenodo.2023997 **?**) , maximum (DOI:https://doi.org/10.5281/zenodo.2559372 **?**) and minimum (DOI:https://doi.org/10.5281/zenodo.2559354 **?**) temperatures

1 Introduction

Long-term observational gridded datasets of near-surface meteorological variables are widely used products. In climatology, they have been used for example to monitor the regional climate (**?**) and to validate and bias-correct climate simulations (**?**). In meteorology, they are used at national meteorological institutes, such as the Norwegian Meteorological Institute (MET Norway), to monitor and report the weather conditions. In hydrology, they are used as external forcing for hydrological and snow modeling (**???**).

seNorge_2018 is a collection of four long-term observational datasets over Norway covering the 61-year time period 1957-2017 for: daily total precipitation (RR), daily mean temperature (TG), daily minimum (TN) and maximum (TX) temperatures. It builds upon the previous work on establishing MET Norway's observational datasets (**???**) and the core of its statistical interpolation method is the Optimal Interpolation (OI, **??**). A review of the relevant literature for our spatial interpolation applications is given in the paper by **?**.

Like the previous versions of seNorge, precipitation and temperature data are provided on a high-resolution grid with 1 km grid spacing in both meridional and zonal directions. seNorge_2018 aims at achieving a higher effective resolution of the analyzed (or predicted) fields than the previous versions. It is worth spending a few words on effective resolution in OI. The difference between grid spacing and resolution is described by **?**. In the context of numerical modeling, **?** defines the "effective resolution as "the minimum wavelength the model can describe with some required level of accuracy (not defined)" and it concludes that as many as 10 gridpoints may be required to properly represent a field. As pointed out by **?**, there is a subjective component in the number of gridpoints needed to resolve a feature in a field. ~~It is worth spending a few more words on effective resolution in OI.~~ In contrast to in-situ observations which represent point values, our gridded analyses produce areal averages. What this means is that for each grid point, we calculate weighted averages of the nearest observations. The ~~extensions of the spatial supports for those averages depend on the settings of the statistical interpolation, which, in turn, are optimized on the station spatial distribution. The~~ larger the extensions of the spatial supports for these averages, the lower the effective resolution of the analysis fields. ~~As demonstrated in the Appendix C of **?**, an OI scheme is realizing a low-pass filter whose cut-off wavelength is determined by the OI settings.~~

~~The main original aspect of our research is that the spatial interpolation methods automatically adapt OI settings to the local station density, such that in data-dense regions the spatial supports of the areal-averaged analyses are smaller than in data-sparse regions. In other words, the effective resolution of~~ In short, it is the ~~analysis fields is higher in data-dense than in data-sparse regions. The user of the seNorge_2018 must be aware that (i) the comparison between different sub-regions over the domain is influenced by the respective local station densities, and (ii) variations in the observational network over time will affect temporal trends derived from this dataset. According to **?**, to overcome those limitations a further post-processing of seNorge_2018 would be required so to create a new dataset, less accurate but suitable for the investigation of long-term variations and trends~~ availability of measurements that determines the highest possible effective resolution, irrespective of the chosen grid spacing, with topographic complexity a compounding factor (**?**). The settings used in the interpolation must

consider this limitation, and if the same settings are to be used over the whole area, then the sub-region of lowest station density may dictate the effective resolution of the entire domain.

The following definitions of spatial scales are used in the text. Regional scale coincides with the whole domain. Given the importance of the observational network, at an arbitrary point we refer to scales that are defined with respect to the station distribution in its surroundings. Sub-regional scale (or local scale) defines an area -around the point- that includes dozens of observations (10-100). Small-scale defines an area that includes few observations (1-10). Unresolved scale refers to those spatial scales that are smaller than the average distance between a station and its closest neighbours, such that atmospheric fields could not be properly represented by the observational network.

The main original aspect of our research is that the spatial interpolation methods automatically adapt OI settings to the local station density, such that in data-dense regions the spatial supports of the areal-averaged analyses are smaller than in data-sparse regions. In other words, the effective resolution of the analysis fields is higher in data-dense than in data-sparse regions. Because the spatial analysis depends on station density, the Integral Data Influence (IDI: **??**) has been used as a diagnostic parameter to quantify the effects of station density on the analysis.

The presented research includes several other original aspects. In the case of precipitation, the measurements have been adjusted for the wind-induced under-catch in a way that is consistent with the method proposed by **?**. A multi-scale OI scheme has been implemented on precipitation relative anomalies with respect to a reference field that captures the field variability at unresolved spatial scales. ~~Furthermore, a Box-Cox transformation has been used to get the precipitation relative anomalies into a normal shape as required by OI. The~~ The reference fields are the monthly totals derived from a regional climate simulation with a resolution of 2.5 km. The climate simulation is based on the dynamical downscaling of the global reanalysis ERAInterim and it is available for the time period 2003-2016. In the paper by **?**, it has been demonstrated that the combination of the same model fields with observed data do improve the representation of monthly total precipitation over Norway. **?** proved that the use of a reference field as a ~~first-guess~~ first guess for the precipitation patterns ~~has been proven to be~~ is a successful approach ~~by ?~~ also in the Alps. They found that daily precipitation over the Alpine region is well represented by using the seasonal precipitation mean as a single predictor field in Kriging with external drift. ~~The precipitation observational network is extremely sparse over significant portions of our domain, such as in mountainous regions, where the vast majority of stations are located on the valley floors (?); and in the Arctic region. As a consequence, we have chosen not to use precipitation climatologies derived by observational gridded datasets. Instead the reference is derived from long-term averages calculated from the output of a high-resolution numerical model. We have used a regional climate simulation with a resolution of 2.5 km, based on the dynamical downscaling of the global reanalysis ERAInterim and available for the time period 2003-2016, to derive the monthly reference fields, as this has been proven skillful for such an application (?).~~

In the case of temperature, seNorge_2018 is the first seNorge dataset ~~including~~ that includes daily minimum and maximum temperatures. The availability of these two additional variables allow for the computation of several more indices for climate variability and extremes, such as the ones reported in the paper by **?**. The three temperature variables are treated separately with the same interpolation method. With respect to seNorge2 (**?**), the regional spatial trend of temperature is obtained as the blending of a much larger number of sub-regional trends. The analysis method has been implemented on a gridpoint-by-

3

gridpoint basis ~~so~~ in order to take advantage of a local Kalman gain. ~~As a result, the effective resolution of the analyzed fields is optimized over the domain.~~

The structure of the paper is as follows. Section **??** presents the observational network and the regional climate simulation used as the precipitation reference. Furthermore, IDI is described and discusses in Sec. **??** as for spatial analysis we regard this parameter as one of the basic properties characterizing a station, such as e.g. its altitude or the geographical location. The spatial interpolation methods are described in ~~Section~~Sec. **??**. ~~Finally,~~ An example application for precipitation is presented in Sec. **??**. The features of seNorge_2018 daily temperature fields are very much similar to those displayed in e.g. Figs.(4)-(6) in the paper by **?**, since the grid is the same and the spatial analyses are based on the same principles. For this reason, example applications for temperature are not included. Section **??** presents the ~~results and the~~ validation of seNorge_~~2018. With its intricate coastline and complex terrain, Norway is an excellent region for testing spatial interpolation schemes under challenging conditions. In this sense, the evaluation presented can be useful to infer the performances of the presented methods over other regions as well~~2018, that is largely based on cross-validation (CV) and comparison against seNorge2. Then, the results are discussed in Sec. **??**.

2 Data

2.1 Observations

The in-situ observations are retrieved from MET Norway's climate database and the European Climate Assessment and Dataset (ECA&D, **?**). The spatial domain covers the Norwegian mainland, plus an adjacent strip of land extending into Sweden, Finland and Russia in order to reduce boundary effects along the Norwegian border. The observations have been quality controlled by experienced staff and with the help of automatic procedures, such as the spatial consistency test described by **?**. The variables are defined as following: TG is the 24-hour average between 06:00 UTC of the day, reported as time-stamp and 06:00 UTC of the previous day; RR is the accumulated precipitation over the same time interval as TG, moreover RR data has been corrected for the wind-induced under-catch of the gauges; TX and TN are, respectively, the maximum and minimum observed temperatures between 18:00 UTC of the day reported as time-stamp and 18:00 UTC of the previous day. TG and RR share the same day-definition so as to ~~better~~ serve hydrological applications ~~, while for historical reasons~~ (**??**). As a result of choices made in the past at MET Norway, TX and TN have a different day definition than RR and TG.

The measured RR value (i.e., RR_{raw}) at an arbitrary location is adjusted for wind-induced under-catch of solid precipitation by means of a procedure similar to the one presented by **?**:

$$\alpha = \tau_1 + (\tau_2 - \tau_1)\left\{\exp\left[(\text{TG} - T_\tau)/s_\tau\right]/\left(1 + \exp\left[(\text{TG} - T_\tau)/s_\tau\right]\right)\right\} \tag{1}$$

$$\gamma = [1 - \alpha]\exp\left[-(W/\theta)^\beta\right] + \alpha \tag{2}$$

$$\text{RR} = \gamma^{-1}\,\text{RR}_{raw} \tag{3}$$

where TG is extracted from the analysis field (Sec. **??**) so to always have a temperature estimate; W is the ten-metre wind speed at the station location extracted from a gridded dataset derived from numerical model output. The (NORA10, **?**) wind speed

dataset, which covers the whole time period 1957-2017, has been downscaled onto the 1 km grid by using a quantile mapping

125 approach (**?**) to match the climatology of the high-resolution numerical weather prediction model (AROME-MEtCoOp, **?**). The wind dataset is available for public download at http://thredds.met.no/thredds/catalog/metusers/klinogrid/KliNoGrid_16. 12/FFMRR-Nor/catalog.html. In the original paper by **?**, they were considering sub-daily precipitation measurements and both temperature and wind were measured at the same location as precipitation. We are operating under different conditions and the requirement of having temperature and wind measurements together with precipitation would reduce the number of

130 suitable observations to a very small subset. As a consequence, in Eqs. (**??**)- (**??**) we had to use parameter values which are different from those used by **?**. We have decided to use seNorge version 1.1 (**??**) as a reference for the extreme values returned by the precipitation adjustment. seNorge version 1.1 includes a precipitation correction based on geographical parameters, summarized in site exposure classes such that a systematic increase of precipitation is carried out. The correction presented in Eqs. (**??**)- (**??**) takes advantage of wind and temperature estimates but we do not expect the extreme values of those two

135 corrections to differ significantly. The parameter values used in Eqs. (**??**)- (**??**), which have been optimized to better match seNorge version 1.1 extremes, are: $\theta = 4.7449$, $\beta = 0.6667$, $\tau_1 = 0.4930$, $\tau_2 = 0.9134$, $T_\tau = 0.9134$, $s_\tau = 0.7759$.

Figure **??** shows the observational network and its evolution in time. The number of available observations was rather stable from 1957 to 2000. In the following decade, the number of RR observations dropped to 500, which was the minimum value, and then it gradually increased again to over 600 in ~~the~~ recent years. The number of temperature observations has been constantly

140 increasing since year 2000, and for 2017 there are about twice as many stations as in 1957. The meteorological stations have been mainly installed to monitor the weather in cities and villages, so the network is denser in urban areas. In the mountainous regions, the digital elevation model (resolution of 1 km^2) can reach 2000 m but most of the stations are located below the elevation of 1000 m. A difference in the station density between the southern and the northern ~~portion~~ portions of the domain is also clearly visible, with a higher density in the south of Norway. Ideally, spatial interpolation would require a denser network

145 of observations where the variance of ~~a variable~~ the field is larger, in order to get a fine-scale representation of the ~~temperature~~ field where it varies the most. However, this is hardly the case in most situations because of the inherent difficulties in station installation and maintenance over complex terrain and in remote areas. As a result, we should expect better performances of the interpolation methods over urban areas and larger analysis uncertainties over data-sparse areas, such as mountainous regions.

~~As stated in the Introduction, the spatial interpolation depends on the station density. For this reason, the Integral Data~~
150 ~~Influence (IDI: **??**) has been introduced as a diagnostic parameter and it is shown in~~

2.2 Reference fields for spatial interpolation of precipitation

The reference fields are derived from long-term averages calculated from the output of a high-resolution numerical model. The reference datasets used for precipitation are based on hourly precipitation provided by the climate model version of HARMONIE (version cy38h1.2), a seamless NWP model framework developed and used by several national meteorological

155 services. HARMONIE includes a set of different physics packages adapted for different horizontal resolutions. For ~~the bottom panels of~~ high-resolution, convection permitting simulations in this case, the model has been set-up with AROME physics (**?**) and the SURFEX surface scheme (**?**). The climate runs have been carried out within the HARMONIE script system, covering

5

the period July 2003 to December 2016 on a 2.5 km grid over the Norwegian mainland. More details on the climate model can be found in **?**, references therein and on https://www.hirlam.org/trac/wiki/HarmonieClimate. The numerical model does

160   not include measurements from the network of rain-gauges. The mean monthly total precipitation fields have been computed considering the available hourly data and they have been used as reference fields for the spatial interpolation of precipitation as described in Sec. **??**.

     Over our domain, we have chosen not to use precipitation climatologies derived by observational gridded datasets as the reference because in some regions the observational network is extremely sparse (Fig. **??**).

165   **2.3**   **Integral Data Influence**

     IDI is similar to the degrees of freedom introduced by **?** and it has been used also to evaluate the distribution of weather stations (**?**). In practice, IDI is obtained as the result of an OI performed by arbitrarily assigning ~~the~~ a value of 1 to the observations (i.e., maximum amount of available information) and the reference value of 0 to the background (i.e, basic amount of information available everywhere). The analytical function that usually represents the background error correlation in OI, in the case of

170   IDI is representing the station influence on the analysis according to a predefined metric. This metric is defined as a function of the geographical parameters. For an arbitrary point in space, the geographical parameters are stored in a vector $\mathbf{r}$ having four components: latitude, longitude, altitude and land area fraction (i.e. fraction of land in the 1 km square box centered at the point). The land area fraction is introduced here and used in Sec. **????**. Functions are applied to pair of points, such as: $d(\mathbf{r}, \mathbf{s})$ returns the horizontal (radial) distance in km between $\mathbf{r}$ and $\mathbf{s}$; $z(\mathbf{r}, \mathbf{s})$ returns their absolute elevation difference; $w(\mathbf{r}, \mathbf{s})$

175   returns their absolute land area fraction difference. The correlation function between two points $\mathbf{r}$ and $\mathbf{s}$ is based on Gaussian functions of the form:

$$f_u(\mathbf{r}, \mathbf{s}; D) = \exp\left\{ -\frac{1}{2} \left[\frac{u(\mathbf{r}, \mathbf{s})}{D} \right]^2 \right\} \tag{4}$$

where: $u()$ is an arbitrary function, such as the ones previously defined, applied to the points; D is a reference length scale governing the decreasing rate. We have chosen to model the station influence using Gaussian functions. For TG, TX and

180   TN, the station influence is factorized into the product of two Gaussian functions: one depending on distances, such that in Eq. (**??**) $u = d()$ and $D =$50 km; the other depending on elevation differences, with $u = z()$ and $D =$200 km. In the case of RR, the station influence depends only on distances, therefore $u = d()$ and $D =$10 km. The values of the de-correlation length scales used for temperature are consistent with the findings of Sec. **??**~~, while for precipitation~~. For precipitation, the value chosen is representative of the smallest spatial scales used in ~~the iterative cycle. Elevation plays always a predominant role~~

185   ~~for temperature and even only a few stations at higher elevations can provide a reasonable approximation of the sub-regional near-surface temperature lapse rate. Then, it is important to know where this information is not available and the temperature IDI map in Fig. ?? shows those regions. On the other hand, for precipitation we have decided to not consider elevation because we are aware that our network is very sparse at higher elevations and for this reason we have introduced a reference field, as stated in the Introduction. The precipitation IDI map in Fig. ?? highlights the potential of our stations to interact on the~~

6

190 ~~horizontal plane. In the following, the cross-validation (CV) approach is used such that the summary statistics derived at station locations can be considered valid for gridpoints. The~~ multi-scale OI of Sec. **??**.

For the purpose of evaluation in Sec. **??**, the CV-IDI at station locations (i.e., IDI at a station location computed without considering the presence of that station) is introduced to link the CV statistics to the IDI of the hypothetical gridpoint represented by a station location.

195 In the two maps of Fig. ~~**??**~~**??**, the IDI is shown ~~over the 1 km grid~~ for TG and RR based on the station distributions shown in Fig. **??**. The IDIs for TX and TN are very similar to ~~the TGmap~~TG. In the vicinity of an observation the IDI field tend to stay close to 1 whereas for data sparse areas its value is close to 0. The IDI and CV-IDI values have been ~~divided into~~ arbitrarily divided into four classes: values smaller than 0.45 ~~defines~~ define observations/gridpoints in data-sparse regions (i.e., where the station influence on the analysis is very limited); values larger than 0.85 ~~defines~~ define observations/gridpoints in data-dense

200 regions (i.e., where the station influence on the analysis is substantial), then two ~~intermediate classes~~ transition classes between data-dense and -sparse regions have been defined~~to have an idea of the behaviour of the spatial interpolation scheme in the transition between data-dense and data-sparse regions~~.

For temperature, elevation plays a predominant role and even only a few stations at higher elevations can provide a reasonable approximation of the sub-regional near-surface temperature lapse rate. Fig. **??** shows that the regions where our observational

205 network is sparser are the Northernmost part of Norway (i.e., above latitude 69 N) and the Scandinavia mountains between latitude 66-68 N. For precipitation, we have decided to not consider elevation in the spatial analysis because we are aware that our network is very sparse at higher elevations (see Fig. **??**).

For precipitation, the IDI map in Fig. **??** shows values larger than 0.85 for those regions where the observational network can reconstruct patterns in the analysis fields where the small-scales have a resolution of approximately 10 km. The largest

210 continuous regions with IDI larger than 0.85 are located in the southern part of the domain (i.e, below latitude 65 N) and mostly along the coast.

Fig. ~~**??**~~**??**a and Fig. ~~**??**~~**??**b show the close ~~relationship~~ relationships between CV-IDI and the station density. In Fig. ~~**??**~~**??**a the distribution of CV-IDI values is shown by boxplots~~: the~~, where each gray box represents a sample distribution: the black thick horizontal line is the median~~of the distribution, the~~; the gray box width is the interquartile range; the whiskers extend to

215 the tails. As shown by Fig. ~~**??**~~**??**c, at station locations ~~the IDI values are confined into~~ IDI has a smaller range of values than ~~the~~ CV-IDI. ~~Even~~ In fact, even an isolated station constitutes more information than the background alone, while an isolated gridpoint must have IDI equals to 0 as it is CV-IDI at an isolated stations.

~~The reference datasets used for precipitation are based on hourly precipitation provided by the climate model version of HARMONIE (version cy38h1.2), a seamless NWP model framework developed and used by several national meteorological~~

220 ~~services. HARMONIE includes a set of different physics packages adapted for different horizontal resolutions. For the high-resolution, convection permitting simulations in this case, the model has been set-up with AROME physics (?) and the SURFEX surface scheme (?). The climate runs have been carried out within the HARMONIE script system, covering the period July 2003 to December 2016 on a 2.5 km grid over the Norwegian main land. More details on the climate model can be found in ?, references therein and on . The numerical model does not include measurements from the network of rain-gauges. The mean monthly~~

225 ~~total precipitation fields have been computed considering the available hourly data and they have been used as reference fields for the spatial interpolation of precipitation as described in Sec. ??.~~

3 ~~Methods~~ Spatial Interpolation methods

The notation used is based on both **?** and **?**. The number of gridpoints is m. The number of observations is p. Upper-case bold symbols are used for matrices, lower-case bold symbols for vectors and italic symbols for scalars. For an arbitrary matrix

230 $\mathbf{X}$, $\mathbf{X}_i$ means the ith column; $\mathbf{X}_{i,:}$ the ith row; and $\mathbf{X}_{ij}$ the element at the ith row and jth column. For an arbitrary vector $\mathbf{x}$, $\mathbf{x}_i$ denotes the ith element. The superscripts on the upper left hand corner of a symbol identify: analysis a; background b; observation o. Upper accents have been used too. In the case of temperature, where we iterate over the gridpoints, the notation $\overset{i}{\mathbf{X}}$ indicates that matrix $\mathbf{X}$ is valid for the ith gridpoint and in this sense we may refer to it as a local matrix. In the case of precipitation, where we iterate over spatial scales, ~~the same notation~~ $\overset{i}{\mathbf{X}}$ those length scales are indicated with greek letters and

235 the notation $\overset{\alpha}{\mathbf{X}}$ indicates that matrix $\mathbf{X}$ is obtained as a function of the ~~ith spatial scale~~ spatial scale of α km. Upper accents are not used only for matrices, for instance the in-situ observations are stored in the p-vector $\mathbf{y}^o$ but in the following we will refer to the $\overset{i}{p}$-vector $\overset{i}{\mathbf{y}}^o$ of the nearest observations to the ith gridpoint.

3.1 ~~Temperature~~ Statistical interpolation of temperature

The same interpolation scheme is used for the mean, the maximum and the minimum daily temperature. The physical con-

240 sistency among ~~them~~ the three variables is assured by post-processing the independently analyzed datasets ~~, such that~~ and for each gridpoint ~~: minimum temperature~~ we impose that: TN is always smaller or equal to ~~the mean, maximum temperature is always bigger~~ TG; TX is always greater or equal to ~~the mean~~ TG. The cross-checking is further discussed in Sec. **??**.

The spatial interpolation is implemented on a gridpoint-by-gridpoint basis. It combines a regional pseudo-background field, that is the weighted average of numerous sub-regional fields, with the observations. The temperature analysis at the generic ith

245 gridpoint is written as:

$$\mathbf{x}_i^a = \mathbf{x}_i^b + \overset{i}{\mathbf{K}}_{i,:} \left(\overset{i}{\mathbf{y}}^o - \overset{i}{\mathbf{y}}^b \right) \tag{5}$$

$\overset{i}{\mathbf{y}}^o$ and $\overset{i}{\mathbf{y}}^b$ are ~~the~~ $\overset{i}{p}$-vectors of the nearest stations to the ith gridpoint.

The local Kalman gain in Eq. (**??**) is:

$$\overset{i}{\mathbf{K}}_{i,:} = \overset{i}{\mathbf{G}}_{i,:} \left(\overset{i}{\mathbf{S}} + \varepsilon^2 \overset{i}{\mathbf{I}} \right)^{-1} \tag{6}$$

250 $\overset{i}{\mathbf{I}}$ is the $\overset{i}{p} \mathbf{x} \overset{i}{p}$ identity matrix and $\varepsilon^2 \equiv \sigma_o^2 / \sigma_b^2$ is the ratio between the constant observed (σ_o^2) and pseudo-background (σ_b^2) error variances that has been set to 0.5, as for seNorge2 (**?**). The local pseudo-background error correlation matrices are defined on the basis of the correlation function between pair of points $\rho^T(\mathbf{r}_j, \mathbf{r}_k)$ as:

$$\rho^T(\mathbf{r}_j, \mathbf{r}_k) = f_d\left(\mathbf{r}_j, \mathbf{r}_k; D_i^h\right) \ f_z\left(\mathbf{r}_j, \mathbf{r}_k; D^z\right) \ [1 - (1 - w_{min})|w(\mathbf{r}_j, \mathbf{r}_k)|] \tag{7}$$

8

such that the correlation between the jth ~~gridgpoint~~ gridpoint and the kth station is $\overset{i}{\mathbf{G}}_{jk} = \rho^T(\mathbf{r}_j, \mathbf{r}_k)$. Analogously, the correlation between the jth station and the kth station is $\overset{i}{\mathbf{S}}_{jk} = \rho^T(\mathbf{r}_j, \mathbf{r}_k)$. The Gaussian functions f are defined in Eq. (**??**). A formulation similar to Eq. (**??**) has been used in the paper by **?**, in that case the land area fraction has been replaced by the land use. w_{min} sets the minimum value for the factor related to land area fraction when $w(\mathbf{r}_i, \mathbf{r}_j)$ is maximum (i.e., equals to 1). D^z and w_{min} are fixed over the domain, while D_i^h is allowed to vary between gridpoints, although with some restrictions. In an ideal situation of a very dense observational network, one may consider to rely on adaptive estimates for the three parameters. This is not the case for our station distribution, so we have opted for a "hybrid" configuration (i.e., D^z and w_{min} fixed; D^h adaptive) that would return robust estimates. The impact of large land area fraction differences on ρ^T is less dramatic than those of large horizontal or elevation differences and it also impacts only a limited number of stations along the coast. Eventually, we have manually set $w_{min} = 0.5$ to achieve the desired effect of attenuating the influence of coastal areas over inland areas and vice versa, while at the same time avoiding the introduction of sharp gradients between those two regions. The optimization procedure for D_i^h and D^z is described in the following of this section.

The pseudo-background x_i^b in Eq. (**??**) is the blending of n sub-regional pseudo-backgrounds and it is in many ways similar to those described by **?**. Each sub-regional pseudo-background is defined by a centroid and it includes only the 30 stations closest to this centroid. The pseudo-background field with centroid at $\mathbf{r}_c$ is the m-vector $\overset{c}{\mathbf{x}}{}^b$ and its value at the ith gridpoint is $\overset{c}{\mathbf{x}}{}^b_i$. The seNorge_2018 domain has been divided on a 50x50 grid, each cell is a 24 km by 31 km rectangular box and the nodes (i.e., centres of the cells) are the "candidate" centroids. If a node is inside the domain and has at least 30 stations in a neighbourhood of 250 km, then it is a suitable centroid. Those 30 temperature observations are used to estimate a sub-regional pseudo-background field as a function of the elevation only. The analytical function used to model the vertical profile of temperature is the one proposed by **?** for the alpine region and its parameters have been obtained by fitting the function to the aforementioned 30 observations. We assume that 30 observations can provide a reliable fitting. The generic cth pseudo-background field $\overset{c}{\mathbf{x}}{}^b$ is derived directly from the digital elevation model by assuming that the cth sub-regional vertical temperature profile is valid for the whole domain. ~~The~~ By using a 50x50 grid, the number of sub-regions n is usually between 500 and 600 and there are significant overlaps between neighbouring sub-regions, such that the continuity of the regional pseudo-background is guaranteed. Finally, $\mathbf{x}_i^b$ is a weighted average of n values:

$$\mathbf{x}_i^b = \frac{\sum_{c=1}^{n} \overset{c}{\mathbf{x}}{}^w_i \, \overset{c}{\mathbf{x}}{}^b_i}{\sum_{c=1}^{n} \overset{c}{\mathbf{x}}{}^w_i} \tag{8}$$

where the weights at the ith gridpoint $\overset{c}{\mathbf{x}}{}^w_i$ are the n IDI values (Sec. **??**) and $\overset{c}{\mathbf{x}}{}^w_i$ is computed considering only those stations included in cth sub-regional pseudo-background field. The settings used in the IDI calculation are similar to those used for precipitation in Fig. **??** **??**, in the sense that the station influence decays with horizontal distance only and its de-correlation length scale is set to 27.5 km, that is the average of a box width and height on the 50x50 grid.

The optimization of D^z and D_i^h of Eq. (**??**) is based on the statistics of the innovation (i.e. observation minus background) at station locations. As described by **?**, the elements of the background error covariance matrix at station locations, which

is modeled by us as $\sigma_b^2 \mathbf{S}$, should match the innovation sample covariances. In Tables **??-??**, the values of the parameters determining $\sigma_b^2 \mathbf{S}$ are shown for a selection of year (1960, 1970, ~~....~~1980, 1990, 2000, 2010) in the assumption of a constant D^h (i.e., $D_i^h = Dh$, $\forall i = 1, \ldots, m$). Note that we have also added the average number of stations available for a specific year, the average distances between them and the estimated observation error variance, which is not strictly required to compute $\mathbf{S}$ and it is set to be half of σ_b^2 in our analysis. The TN error variances are significantly higher than those for TG and TX, thus indicating that TN is a more challenging variable to interpolate. D^h and D^z do not differ significantly among TG, TX and TN, probably because the common observational network ~~is the main constrain~~ constitutes the major constraint in determining their value. This justifies our choice to set $D^z = 210$ m for the three variables. The parameter values in the Tables are more influenced by the majority of stations that are located in station-dense areas. Therefore, the value of $D^h = 55$ km can be considered as a suitable reference for the minimum allowed D_i^h value. The procedure used for the D_i^h estimates is similar to the one described for the regional pseudo-background field. D_i^h is a weighted average as the one reported in Eq. (**??**) where $\overset{c\,b}{\mathbf{x}}_i$ is replaced by the cth length scale, which is constant for all gridpoints. For the cth sub-region this length scale is set to the average distance between a station and its nearest ~~4~~3 stations, provided that this distance is larger than $D^h = 55$ km, otherwise $D^h = 55$ km is used. In this way, the analysis in data-sparse regions is the result of an interaction between a few (approximately four) stations. At the same time, we take advantage of data-dense areas to increase locally the effective resolution of the analysis without destroying the continuity of the field. Note that the use of extremely different D^h values between data-dense and sparse areas (i.e., with differences around one order of magnitude or more) would result in a rather confusing field to look at. In those cases it would be better to split the domain ~~in~~ into sub-domains and operate independently on them.

3.2 ~~Precipitation~~Statistical interpolation of precipitation

The multi-scale OI analyses are the results of successive approximations of the observations over a sequence of decreasing spatial scales that at station locations converge to the observed values.

The interpolation scheme is not applied directly to the RR values (the vector of the raw observed values adjusted for the wind-induced under-catch is indicated as $\mathbf{y}^{\mathrm{rr}}$) but to their anomalies relative to a reference field of monthly precipitation (see Sec. **??**, indicated with the abbreviation ref in the following). In addition, a Box-Cox transformation with power parameter set to 0.5 is used and the transformation is indicated with the function $g()$. A similar transformation has been suggested by **?**, though in the context of combination of radar with gauge data. The ith element of $\mathbf{y}^o$ used in multi-scale OI is:

$$\mathbf{y}_i^o = g\left(\mathbf{y}_i^{\mathrm{rr}} / \mathbf{y}_i^{\mathrm{ref}}\right) \tag{9}$$

The analysis procedure can be written as:

$$\mathbf{x}^a = g^{-1}\left[\mathcal{M}_2 \circ \mathcal{M}_3 \circ \ldots \circ \mathcal{M}_\eta (\tilde{\mathbf{x}}^a)\right] \odot \mathbf{x}^{\mathrm{ref}} \tag{10}$$

where the three fundamental operations are: (1) the composition of several applications of the same statistical interpolation model down a hierarchy of spatial scales of ~~{l km, ..., 3 km, 2 km}~~{η km, ..., 3 km, 2 km}, such that the results of a model application are used to initialize the successive one, $\circ$ stands for model composition and ~~$\mathcal{M}_t$~~$\mathcal{M}_\eta$ stands for the application of

the statistical model to the ~~largest~~ length scale of ~~l km, besides $\overset{\eta}{\mathbf{x}}^a$~~ η km, $\tilde{\mathbf{x}}^a$ is the average of the $\mathbf{y}^o$ elements; (2) the Box-Cox inverse-transformation $g^{-1}()$; (3) the elementwise multiplication of vectors $\odot$ so to transform the relative anomalies into RR values and at the same time include the effects of unresolved spatial scales. Ideally, the sequence of spatial scales to be used in Eq. (**??**) should be bounded between a very large scale (e.g, half the largest domain dimension) and a fine scale corresponding to the average distance between two stations in data-dense areas. The number of scales in between those two extremes is not critical for the final results, provided that they are enough to guarantee a continuous analysis field in all situations. For seNorge_2018, we are using approximately 100 scales with a minimum of 2 km and a maximum of 1400 km. According to **??**, those spatial scales range from the regional synoptic down to the lower boundary of the mesoscale. The sequence is unevenly spaced as the difference between two consecutive scales is somewhat proportional to their values.

The step-by-step description of the model ~~$\overset{\alpha}{\mathbf{x}}^a = \mathcal{M}_s\left(\overset{\beta}{\mathbf{x}}^a\right)$ for the arbitrarily length scale of s km~~ $\overset{\alpha}{\mathbf{x}}^a = \mathcal{M}_\alpha\left(\overset{\beta}{\mathbf{x}}^a\right)$ for two successive and arbitrary length scales of β km and α km (with $\beta > \alpha$) is:

$$\overset{\alpha}{\mathbf{x}}^b = \overset{\alpha}{\boldsymbol{\Psi}}\overset{\beta}{\mathbf{x}}^a \tag{11}$$

$$\overset{\alpha}{\mathbf{K}} = \overset{\alpha}{\mathbf{G}}\left(\overset{\alpha}{\mathbf{S}} + \varepsilon^2 \overset{\alpha}{\mathbf{I}}\right)^{-1} \tag{12}$$

$$\overset{\alpha}{\mathbf{x}}^a = \overset{\alpha}{\mathbf{x}}^b + \overset{\alpha}{\mathbf{K}}\left(\mathbf{y}^o - \overset{\alpha}{\mathbf{H}}\overset{\alpha}{\mathbf{x}}^b\right) \tag{13}$$

In order to reduce the multi-scale OI computational expenses, the original 1 km grid is aggregated onto a new coarser grid, with aggregation factor equal to the integer value nearest to ~~$s/2$~~ $(\alpha/2)$ km. The aggregation groups several smaller rectangular boxes into a bigger one (e.g., if ~~$s=8$~~ $\alpha = 8$ km then 16 of the 1 km by 1 km boxes are aggregated into a single box measuring 4 km by 4 km). The analysis at scale ~~$s+1$~~ β is used as the background for scale ~~s~~ α in an OI scheme. The analysis values are transferred between the two non-matching grid by the operator $\overset{\alpha}{\boldsymbol{\Psi}}$ of Eq. (**??**), that is a bilinear interpolation mapping vectors on the ~~$(s+1)$~~(coarser) β-grid to vectors on the ~~s~~(finer) α-grid. The observation operator $\overset{\alpha}{\mathbf{H}}$ of Eq. (**??**) is also a bilinear interpolation transforming vectors on the ~~s~~α-grid to ~~observation vectors~~p-vectors. ε^2 is set 1, which means that observations and background are assumed the have the same error variances. The background error correlation matrices of Eq. (**??**) are defined on the basis of the correlation function ρ^R:

$$\rho^R(\mathbf{r}_j, \mathbf{r}_k) = f_d\left(\mathbf{r}_j, \mathbf{r}_k; \underline{s}\alpha\right) \tag{14}$$

The Gaussian function f is defined in Eq. (**??**) and the notation is similar to Eq. (**??**). Note that the length scales enter multi-scale OI of Eq. (**??**) through the correlation function of Eq. (**??**). An OI scheme such as the one presented in Eq. (**??**)- (**??**) is realizing a low-pass filter whose cut-off wavelength is approximately ~~s~~α km (**?**) so every iteration over a smaller spatial scale returns a field with more fine-scale details in it. For a given element of the observation vector, there may be a "critical" scale at which the background coincides exactly with the observed value such that its contribution to the innovation in Eq. (**??**) (i.e., the term in ~~parenthesis~~parentheses) is equal to zero and its analysis value would not change over the subsequent iterations. That critical scale is variable across the domain and depends on both the spatial structure of the precipitation field and the local station density.

4 ~~Results~~Example application for precipitation

An example of application of the spatial interpolation method described in Sec. **??** is given in this section. We have chosen to illustrate the analysis procedure applied to the day 1998-10-24, that is one of the days within the period 1957-2017 with the highest value of averaged observed precipitation and where almost all the gauges have measured precipitation.

The reference field for October (see Sec. **??**) is the prior information used in our spatial analysis and it is shown in Fig. **??**. The highest precipitation values occur in the western part of Norway and a precipitation gradient between the coast and inland areas is evident.

The first of the three fundamental steps in Eq. (**??**) is the iterative application of OI for the Box-Cox transformed relative anomalies over a sequence of decreasing spatial scales. We are looping over 91 scales and in Fig. **??** three of those scales are shown. As a result of the transformation, the fields on the maps are not straightforward to interpret as they do not correspond to e.g. precipitation. For this reason, we have chosen an ad-hoc colour scale that highlights the patterns in the fields more than the differences among values. For each scale, the OI is implemented as in Eqs. (**??**)- (**??**). The smaller the spatial scale, the more fine-scale details the OI will represent and, as a consequence, the finer is the grid used. For the scales of 100 km, 31 km and 2 km, the interpolation are performed over grids of 50 km, 15 km and 1 km, respectively. From the coarser scale represented in Fig. **??**, it is only possible to have a rough idea of the final field as this is only the 17th iteration of the 91 totals. At the scale of 31 km, that is iteration 69, the main features of the field are easier to recognize: the largest values are in the eastern part of the domain, where the reference values are smaller; in the middle of Southern Norway, the field reaches its minimum. The smallest scale is equal to 2 km. With reference to the station distribution in Fig. **??**, the patterns become smaller in data-dense regions while the largest "blobs" occur in data-sparse regions. As one may expect, the patterns on the 1 km grid in Fig. **??** matches the spatial structure of the patterns in Fig. **??**.

Figure **??** is related to the multi-scale OI represented in Fig. **??**. For each gridpoint, the field shows the spatial scale where the last significant variation of the interpolated precipitation relative anomaly has occurred. This scale can be considered an estimate at gridpoints of the critical scale mentioned in Sec. **??**. In particular, it is interesting to know where the precipitation field represents features at the smallest spatial scales included in the multi-scale OI (i.e., from 2 km to 20 km). Those spatial scales correspond to the meso-γ atmospheric scale and are suitable to proper represent e.g., thunderstorms (**?**). It is less relevant to distinguish between spatial scales above 30 km and up to 200 km as they belong to the same meso-β scale e.g, fronts, thunderstorm groups. Fig. **??** emphasizes the gradient between meso-γ and meso-β scales. Where the observational network is dense, the estimated critical scale can assume a wide range of values, that is from 2 km to 100 km, depending on the local characteristics of precipitation (e.g., stratiform or convective). On the other hand, in the data-sparse regions marked as red in Fig. **??**, the station network poses a constraint on the smallest scales the analysis can represent and, as a consequence, the critical scales belong to the meso-β scale, regardless of the precipitation type.

Fig. **??** shows the RR analysis field as derived from Eq.(**??**). It is a combination of the reference field shown in Fig. **??** and the multi-scale OI field shown in Fig. **??** for the scale of 2 km. In particular, the local variations of precipitation in almost data-void regions (e.g, some mountainous areas) are mostly due to the reference field.

5 Verification

The evaluation is based mostly on CV exercises and comparison against the seNorge2 datasets of RR and TG. The cross-validation analysis (i.e., CV-analysis) is the analysis value at a station location obtained considering a selection of the available observations that does not include the one measured at that location. If CV is applied systematically to all stations and it includes all the remaining observations then it is called leave-one-out cross-validation (LOOCV).

The summary statistics of the following variables are used: CV-analysis residuals (i.e, CV-analyses minus observations); innovations (i.e., background minus observations); and analysis residuals (i.e., analyses minus background). The CV-analysis, background and analysis are evaluated through the statistics of CV-analysis residuals, innovations and analysis residuals, respectively. Note that the background is not considered in the verification of precipitation. At a generic station location, CV-analysis and background are independent from the observation, while the analysis has been computed using the observation. As a consequence, the statistics of CV-analysis residuals and innovations have similar interpretations. The CV-analysis residual distributions are used in place of the unknown analysis error distributions at gridpoints. The innovation distributions are used to investigate the properties of the background error at gridpoints. On the other hand, the statistics of analysis residuals reveal the filtering properties of the statistical interpolation at station locations that are related to the observation representativeness error (???). The mean absolute error (MAE) and the root mean square error (RMSE) quantify the average mean absolute deviation and the spread, respectively, of a variable.

For temperature, we are using LOOCV. The comparison between the statistics of CV-analysis residuals and innovations quantifies the improvement of the analysis over the pseudo-background at gridpoints. The fraction of errors (i.e., absolute deviations) greater than $3°C$ has been used as a measure of the tails of the distribution of deviation values. Note that the threshold of $3°C$ is used in MET Norway's verification practice to define a significant deviation that undermines the user's confidence in the forecast.

For precipitation, LOOCV is computationally too expensive. Thus, for each day a random sample of 10% of the available stations have been reserved for CV and not used in the interpolation.

Because precipitation errors follow a multiplicative rather than an additive error model (?), large errors for precipitation are defined as absolute deviations between CV-analysis and observation greater than 50% of the observed value.

5.1 Temperature

The main factors determining the quality of the temperature datasets are: the season of the year, the station density and the terrain complexity. We have also investigated the variations of the performances of our interpolation scheme between two

~~different time periods, 1961-1990 (61-90) and 1991-2015 (91-15), and the evaluation scores are similar to the ones presented in the following for the whole 61-year time period.~~

5.1.1 Summary statistics of the verification scores at station locations

In Figures **??**-**??** the TG, TX and TN evaluation results are shown for summer and winter, respectively, as a function of CV-IDI
420 (Sec. ~~**??**~~**??**). The ~~background, analysis and CV-analysis are evaluated through the statistics of innovation, analysis residuals and CV-analysis residuals, respectively.~~ whole 61-year time period is considered. For each of the four ~~CV-ID~~ CV-IDI classes (Sec. ~~**??**~~**??**), the mean value of the score is displayed. The CV-analysis and background always score better in data-dense areas, as expected. ~~On the contrary, the analysis may show the best results in data-sparse regions because an isolated observation constrains the analysis to fit its value almost exactly.~~ The analysis evaluation scores do not vary much with respect to variations
425 in station density, thus indicating that the observation representativeness error is rather constant for all stations. This result support the use of a single value characterizing the whole observational network for ε^2 in Eq. (**??**).

For all variables, the spatial interpolation scheme generally performs better during summer than winter when small-scale processes (e.g., strong temperature inversion) are more frequent. The TG analysis error distribution at gridpoints, as estimated by CV-analysis residuals, shows that ~~: during~~ during the summer, the MAE is between 0.5°C and 1°C and its RMSE is also
430 around 1°C, and errors larger than 3°C are unlikely even in data-sparse regions; during winter, MAE and RMSE double their values and the differences between data-dense and data-sparse regions are larger, with the probability of having large errors ~~is being~~ approximately 25% in data-sparse regions. The TX analysis error ~~behave~~ behaves similarly to TG~~, it is worth remarking~~. Note that the spatial interpolation method during summer performs better on TG than on TX, while in winter the opposite occurs. ~~TN is the most challenging variable to represent, possibly because atmospheric processes at unresolved spatial scales~~
435 ~~occur more frequently for this variable than for the others.~~ The distribution of the TN analysis error at gridpoints has a much larger spread than those of TG and TX~~, the~~. The TN RMSE is ~~:~~ between 1.5°C and 2°C in summer~~;~~, and up to approximately 4°C during winter in data-sparse regions. The tail of the TN distribution is also longer and large errors are more frequent than for the other daily temperatures. At the same time, the average bias (MAE) is also larger ~~. Our results indicate that a denser station network would be needed to deliver TN products having the same quality as TG and TX~~for TN.

440 ### 5.1.2 seNorge_2018 and seNorge2 comparison of TG

The seNorge_2018 spatial interpolation procedure builds upon seNorge2. Several modifications have been made, though keeping the scale-separation approach. In seNorge_2018 a single function has been used to model the sub-regional vertical profile, instead of the three different functions used in seNorge2. At the same time, in seNorge_2018 the blending of sub-regional fields into a regional pseudo-background field is based on a much larger number of sub-regional fields.

445 In Figure **??** the TG dataset is compared with seNorge2 over two multi-year time periods for the winter season. In the other seasons ~~the two datasets are much more similar, the patterns of deviations still follow those of Fig. **??** but~~ (not shown here) the differences are almost always between −2°C and 1°C. The maps in Fig. **??** show the average daily difference between the analyses of the two datasets. The lateral and bottom panels are shown ~~so~~ to give a better overview of the extreme values. For

most of the Norwegian mainland, seNorge_2018 is colder than seNorge2, with the larger differences on the mountain tops and along the coast in the North. The two periods show similar patterns, however for the 1991-2015 period there are some regions where seNorge_2018 is warmer than seNorge2, such as the plateau in the North between Finland, Russia and Norway; along the border between Sweden and Norway; and in the mountains of Southern Norway.

We believe that the variations between seNorge_2018 and seNorge2 having the most significant impacts on the differences shown in Fig. **??** are in the OI settings. seNorge_2018 includes the land area fraction as an additional parameter in Eq. (**??**) and this improvement causes the differences along the coastline, where stations having significantly different land area fractions are less correlated between each other. By considering only stations along the coast and CV-analysis residual data from 1957 to 2017, seNorge2 has MAE=0.79°C and RMSE=1.12°C, while seNorge_2018 has MAE=0.7°C and RMSE=1.03°C. If compared to seNorge2, seNorge_2018 reduces the bias along the coast by 10% and the RMSE by 8%.

The evident differences between the two datasets are in the mountains, where seNorge_2018 often presents warmer valley floors and colder ridges. In particular, according to Fig. **??**, the highest peaks of the Scandinavian mountains are on average up to 9°C colder, while the valley floors are a few degrees warmer. Those differences can be explained by (1) the reduced D^z value in seNorge_2018 ($D^z = 210$ m instead of $D^z = 600$ m as in seNorge2), which narrows the vertical layer where OI adjustments are effective, therefore temperatures in data-sparse high-altitude regions mostly coincides with background values; (2) the modified procedure for the pseudo-background calculation that realizes a smoother transition between neighbouring sub-regional pseudo-background fields. By considering all the stations and CV-analysis residual data from 1957 to 2017, seNorge2 has MAE=0.81°C and RMSE=1.18°C, while seNorge_2018 has MAE=0.76°C and RMSE=1.18°C.

5.2 Precipitation

5.2.1 Summary statistics of the verification scores

In Figure **??**, the RR dataset is evaluated through the statistics of CV-analysis residuals. In general, the spatial interpolation performs significantly better for station-dense areas, then the performances degrade faster for data-sparse regions as shown by the MAE and RMSE for observations greater than 1 mm/day. The ability in distinguishing between precipitation and no-precipitation is shown by the Equitable Threat Score (ETS) with a threshold of 1 mm/day. In station-dense areas the fraction of correct predictions, accounting for hits due to chance, is approximately 70%, while the ETS goes down to 0.4 in data-sparse

15

regions. With respect to intense precipitation (i.e., observed values greater than 10 mm/day), the probability of having large errors is less than 5% in data-dense areas and around 30% in data-sparse areas.

The close relationship between the terrain and the annual total precipitation is shown in Fig. **??**, where values up to 4700 mm/year are reconstructed along the Norwegian coast.

5.2.2 seNorge_2018 and seNorge2 comparison

Figure **??** shows the differences between seNorge_2018 and seNorge2 mean annual total precipitation in the period 1957-2015. On average, seNorge_2018 has significantly higher precipitation values than seNorge2, especially in data-sparse mountainous regions, and this is due to: i) the correction of rain-gauge data for the wind-induced under-catch, that increase the observed precipitation mostly during winter, and ii) the modified statistical interpolation scheme that uses more information in data-sparse regions, which often tends to increase the precipitation analysis when compared to seNorge2.

For precipitation, we have stated in the Introduction that seNorge_2018 uses a reference to increase the effective resolution of the field, compared to the resolution given by the spatial distribution of the observational network alone. In the context of forecast verification, **?** decompose precipitation fields on different spatial scales using a two-dimensional discrete Haar wavelet decomposition. As described in the paper by **?**, this wavelet decomposition can be used to study the average energy per cell (hereafter energy) of the scale components of precipitation, where the energy is defined as "the average of the field gridpoint squared value". Fig. **??** shows the mean energy of the scale components for seNorge_2018 ~~benefits from the latest efforts in data collection and quality control made by MET and~~ and seNorge2 based on ~~the~~ ~~ECA&D team~~25% of cases with the higher values of averaged precipitation over the domain within the time period 1957-2017. The left panel shows the energies, where seNorge_2018 has more energy than seNorge2 for all scales, a result that is in agreement with Fig. **??**. The shapes of the two energy distributions are similar, with a maximum at the scale of 128 km and a gradual decrease of the energy for smaller spatial scales. However, the right panel shows that the energy percentages for seNorge_2018 are systematically larger than for seNorge2 for scales smaller than 128 km. Because the effective resolution of seNorge2 is determined by the observational network only, the results presented in Fig. **??** support our initial statement on the increased effective resolution of seNorge_2018.

5.3 Occurrence of large errors as a function of the station density

In this paragraph we discuss more in detail the relationships between the occurrence of large errors, as they have been defined at the beginning of Sec. **??**, and the observation densities. We have investigated those relationships for both temperature and precipitation aiming at quantifying the impacts on the analysis quality due to variations in the station distribution.

In Figure **??**, the expected percentage of large errors ~~(i.e., analysis minus true temperature larger than 3°C)~~ at gridpoints is shown for ~~TG, TX and TN~~ all variables during wintertime. ~~This information completes the evaluation presented in Figures ??-?? because it shows the extensions on the grid of the regions characterized by the different data densities. The~~ The close relationships between CV-IDI and the occurrence of large errors are evident from the scatter-plots in Fig. **??** (bottom row) and Fig. **??** (panel d). The same relationships are also represented in the scatter-plots in the panels on the bottom left of each ~~maps are different~~

16

515 ~~representations of the data in the bottom row of~~ of the maps in Fig. **??**, ~~where the points represents the percentage of large errors~~ **??**. If we define the occurrence of large error (err~~at station locations~~) as a function of CV-IDI ~~. The lines~~ (x) as $\mathrm{err}(x)$, then the points in the scatter-plots of Fig. **??** represent the actual occurrence of large errors at station locations and the lines are the best fit to the points of the Gaussian function:

$$\mathrm{err}(x) = a \ \exp\left[-0.5\frac{(x-b)^2}{c^2}\right] \tag{15}$$

520 ~~where x is~~ The estimated large error occurrence at gridpoints is obtained by replacing the CV-IDI ~~for the function fitting performed at station locations, then~~ with the IDI fields ~~shown in~~ (Fig. **??** ~~are used as x values to estimate err over the grid~~**??**) in Eq. (**??**). The three parameters of the bell curve shape are: the value of the curve's peak a; the position of the center of the peak b, and c which controls the width of the bell. The Gaussian function provides reasonably good fits to the points, the relationship between the station density and the analysis quality is non-linear and the analysis performances decay much faster

525 in data-sparse than in data-dense regions. The parameters values for the three variables are: TG $a = 703.012$, $b = -2.626$, $c = 1.123$; TX $a = 856.420$, $b = -2.924$, $c = 1.114$; TN $a = 448.39$, $b = -2.891$, $c = 1.431$. The probability of having a large analysis error is remarkably small over the domain for TX, while for TN such large errors are quite common. The situation for TG is somewhat in between those two extreme cases and large regions of the domain are unlikely to present large errors. Once again, it is evident that the worst performances occur in those regions characterized by complex terrain and low station density,

530 such as the mountainous region between Norway and Sweden in the northern part of the domain.

As for temperature, the expected percentage of large errors over the precipitation grid is shown in Fig. **??**. The three parameters of Eq. (**??**) for RR are: $a = 50.438$, $b = -0.676$, $c = 0.846$. The elevation is not considered in the IDI definition for RR, so the map looks rather different than the temperature maps. Furthermore, by setting $D = 10$ km in Eq. (**??**) (instead of $D = 50$ km as for temperature) we have imposed a fast transition between data-dense and data-sparse regions. For data-dense

535 regions, the expected percentage of large analysis errors for intense precipitation is less than 10%. For data-sparse regions, the percentage of large errors is approximately 40%.

5.4 ~~Precipitation~~

~~The main factors determining the quality of the precipitation dataset are:~~

6 Discussion

540 Because the presented statistical interpolation methods automatically adapt to the local observation density, the ~~station density and the terrain complexity. The season of the year seems to have a smaller impact on the verification scores.~~ user of the seNorge_2018 dataset must be aware that: (i) the comparison between different sub-regions over the domain is influenced by the respective local station densities, and (ii) variations in the observational network over time will affect temporal trends derived from this dataset (**?**).

545 ~~In Figure ?? , the RR dataset is evaluated through the statistics of CV-analysis residuals. In general, the spatial interpolation performs significantly better for station-dense areas, then the performances degrade faster for data-sparse regions as shown by the MAE and RMSE for observations greater than 1 mm/day. The ability in distinguishing between precipitation and no-precipitation is shown by the Equitable Threat Score (ETS)with a threshold of 1 mm/day. In data dense areas the fraction of correct predictions, accounting for hits due to chance, is approximately 70%, while the ETS goes down to 0.4~~ For the

550 four variables, we have investigated the variations of the performances of our interpolation schemes between two different time periods, 1961-1990 (61-90) and 1991-2015 (91-15). The evaluation scores are similar to the ones presented in Sec. ?? for the whole 61-year time period, thus indicating that the ongoing climate change does not have a significant impact on the reconstruction skills of our statistical interpolation schemes.

The three main factors determining the quality of the temperature datasets are the season of the year, the station density

555 and the terrain complexity. The last two factors are correlated, as shown by Figs. ?? - ?? . TN is the most challenging variable to represent, possibly because atmospheric processes at unresolved spatial scales occur more frequently for this variable than for the others. It is worth pointing out that Fig. ?? shows that during the summer TN has a smaller representativeness error in data-sparse ~~regions~~than in data-dense regions, even though the differences are rather small. It is not clear whether this is due to the spatial interpolation scheme or to the occurrence of specific atmospheric phenomena (e.~~With respect to intense~~

560 ~~precipitation (i.e., observed values greater than 10 mm/day), the probability of having large errors (i.e, absolute deviations between CV-analysis and observation greater than 50% of the observed value) is less than 5% in data-dense areas and around 30% in data-sparse areas.~~ g, urban heat island effect). In general, our results indicate that a denser station network is needed to deliver TN products having the same quality as TG and TX by using the statistical interpolation method presented in Sec. ?? . Future developments of the method may also improve the representation of TN, such as: (1) consider the differences between

565 TG and TN, instead of TN; (2) develop a relationship for the vertical profile of TN, that would replace the one proposed in Sec. ?? ; (3) use numerical model output fields (e.g., from reanalysis) in addition to the observed data.

~~The close relationship between the terrain and the annual total precipitation is shown in Fig. ?? , where values up to 4700 mm/year are reconstructed along the Norwegian coast. Figure ?? shows also the differences between seNorge_2018 and~~ The two main factors determining the quality of the precipitation dataset are the station density and the terrain complexity. The

570 season of the year seems to have a smaller impact on the verification scores.

With respect to seNorge2~~mean annual total precipitation in the period 1957-2015. On average~~, seNorge_2018 ~~has significantly higher precipitation values than seNorge2, especially in data-sparse mountainous regions, and this is due to: i) the correction of rain-gauge data for the wind-induced under-catch, that increase the observed precipitation mostly during winter, and ii) the modified statistical interpolation scheme that uses more information in data-sparse regions, which often tend to increase the~~

575 ~~precipitation analysis when compared to~~ presents several methodological improvements and two additional variables, TX and TN. Furthermore, it should be mentioned that there had been variations in the observational datasets used for the production of the two gridded datasets. Even though the data sources are the same for both seNorge datasets, seNorge2 ~~. As~~ is based on an observational dataset that has been produced in 2016, while seNorge_2018 benefits from the latest efforts in data collection and quality control made by MET and the ECA&D team. The verification results show that seNorge_2018 temperature predictions

18

580 are on average more accurate than seNorge2, especially along the coast. With respect to the predictions of precipitation, as described in the seNorge2 paper **?**, this dataset is likely to underestimate precipitation so we have designed seNorge_2018 to returns higher precipitation values because this would better agree with the Norwegian water balance and eventually improve the results of hydrological simulations based on seNorge_2018 than for seNorge2. This last point ~~deserves~~ needs to be verified in the near future.

585 ~~As for temperature in~~ In Sec. **??**, ~~also for precipitation the expected percentage of large errors over the grid is shown in Fig. ??. The three parameters of Eq. for RR are: $a = 50.438, b = -0.676, c = 0.846$. The elevation is not considered in the RR IDI definition, so the map look rather different from the temperature maps. For data-dense regions,~~ **??**, it has been demonstrated that seNorge_2018 has an higher effective resolution than seNorge2. In conclusion, we recommend the use of seNorge_2018 instead of seNorge2 for both TG and RR.

6.1 TG, TX and TN cross-checking

The cross-checking is mentioned at the beginning of Sec. **??** and it is done on the analyses. TG is used as a reference value for both TX and TN, since: (a) the construction of the pseudo-background field is based on a vertical profile relationship that has been developed for TG; (b) TG is expected to be a more robust variable than TX and TN (in the sense defined by e.g., (**?**)), because it is an averaged value instead of an extreme. For each gridpoint, we check whether TN (TX) is smaller (greater) than
595 TG. Wherever the condition is not satisfied, TN (TX) is replaced by TG.

Because of the 12-hour offset in the definitions of TG and either TX or TN, the ~~expected percentage of large analysis errors for intense precipitation is less than 10%. With respect to~~ cross-checking can be wrong. Nevertheless, it is a useful check to identify those situations where the interpolation of daily extremes is not convincing. In fact, despite the offset in the definitions of the ~~IDI calculation (Sec. ??), by setting $D = 10$ km in Eq. (instead of $D = 50$ km as for temperature) we have imposed a fast~~
600 ~~transition between data-dense and data-sparse regions. For data-sparse regions the percentage of large errors is approximately 40%~~24-hour aggregation period, for a typical day TN is smaller than TG and TX is greater than TG. We have found very rare exceptions to these rules in the surroundings of station locations. In the vast majority of cases, the cross-checking flags those gridpoints in the mountains (i.e., far from station locations) where the extrapolation of the vertical profile cannot be adjusted by means of observed data, which are located in large part on the valley floors (Fig. **??**). As a result, the simulated daily extremes
605 are affected by significant uncertainties and we use TG as a reference value to assess the significance of those uncertainties. When the cross-check reports possible violation of the physical consistency, our choice is to "mask out" the analyses of daily extremes and at the same time we propose TG as an alternative.

The ~~evaluation carried out in Fig. ?? for both temperature and precipitation quantifies impacts on the analysis quality due to variations in the station distribution, thus providing a useful tool in the strategic planning of future observational~~
610 ~~networks~~number of gridpoints flagged by the cross-checking vary seasonally and it is higher in winter and lower in summer. In the case of TN, the cross-checking flags on average 9% of the gridpoints in winter and 1% in summer. In the case of TX, the cross-checking flags on average 7% of the gridpoints in winter and 1% in summer.

Future developments will focus on improving the cross-checking, in order to properly handle those exceptional situations that are currently erroneously flagged as physical inconsistencies among the three variables.

7   Conclusions

seNorge_2018 provides 61-year (1957-2017) datasets of daily mean, maximum, and minimum temperatures, ~~and~~ as well as daily total precipitation, over Norway and parts of Finland, Sweden and Russia. The plan at MET Norway is to update the historical dataset once a year, while at the same time provisional daily estimates for the current year are computed every day. MET Norway has an open data policy and all the datasets, as well as most of the observations used in the calculations, are available for public download via its web services.

The observational datasets have been obtained through statistical methods that build upon our previous works. The interpolation schemes automatically ~~adapts~~ adapt their settings to the local station density and this allows for ~~an~~ a higher effective resolution in data-dense areas, while in data-sparse regions the analysis is always the estimate of at least a few stations.

The main factor determining the quality of the temperature analysis are: the season of the year, the station density and the terrain complexity. In the case of precipitation, those factors are: the station density and the terrain complexity. Because of the importance of the combination of station density and terrain, we have widely used the IDI concept in our evaluation.

The new seNorge_2018 shows significant differences when compared to its predecessor seNorge2, both for TG and especially for RR. While first qualitative evaluations indicate that this is an improvement, an indirect evaluation where seNorge_2018 would be used as the forcing data for snow- and hydrological modeling is needed to confirm this.

seNorge_2018 is ~~the first~~ MET Norway's first observational dataset providing TX and TN from 1957. The temperature analysis has the largest errors during winter and the TN is the most challenging variable to represent. For TG and TX, large analysis errors are expected only in winter and limited to almost data-void areas ~~, such as the~~ such as mountain tops. TN may present large analysis errors more often than TG and TX and for larger portions of the domain, especially in mountainous regions.

~~The long-term average of the RR dataset is shaped by the~~ To fill commonly occurring spatial gaps for RR in data-sparse regions, the interpolation uses monthly fields of a high-resolution numerical model ~~, so to fill the gaps in data-sparse regions that are common, given the high variability of daily precipitation~~ and adjusts this to an optimal fit with the measurements that are available in the area. As a result, seNorge_2018 has a finer effective resolution than seNorge2. The ability of the method to correctly distinguish between precipitation and no-precipitation depends critically on the station density. In the North, the sparser observational network is associated with a high occurrence of large analysis errors. The evaluation shows that large analysis errors are unlikely in the data-dense regions of Southern Norway, even for intense precipitation.

8   Code and data availability

The spatial interpolation software is available at (DOI:https://doi.org/10.5281/zenodo.2022479, **?**).

The open-access datasets are available for public download at:

645   – daily total precipitation (DOI:https://doi.org/10.5281/zenodo.2082320, **?**)

     – daily mean temperature (DOI:https://doi.org/10.5281/zenodo.2023997, **?**)

     – daily maximum temperature (DOI:https://doi.org/10.5281/zenodo.2559372, **?**)

     – daily minimum temperature (DOI:https://doi.org/10.5281/zenodo.2559354, **?**)

seNorge_2018 is daily updated by MET Norway and the most recent data are available at http://thredds.met.no/thredds/catalog/
650   senorge/seNorge_2018/catalog.html.
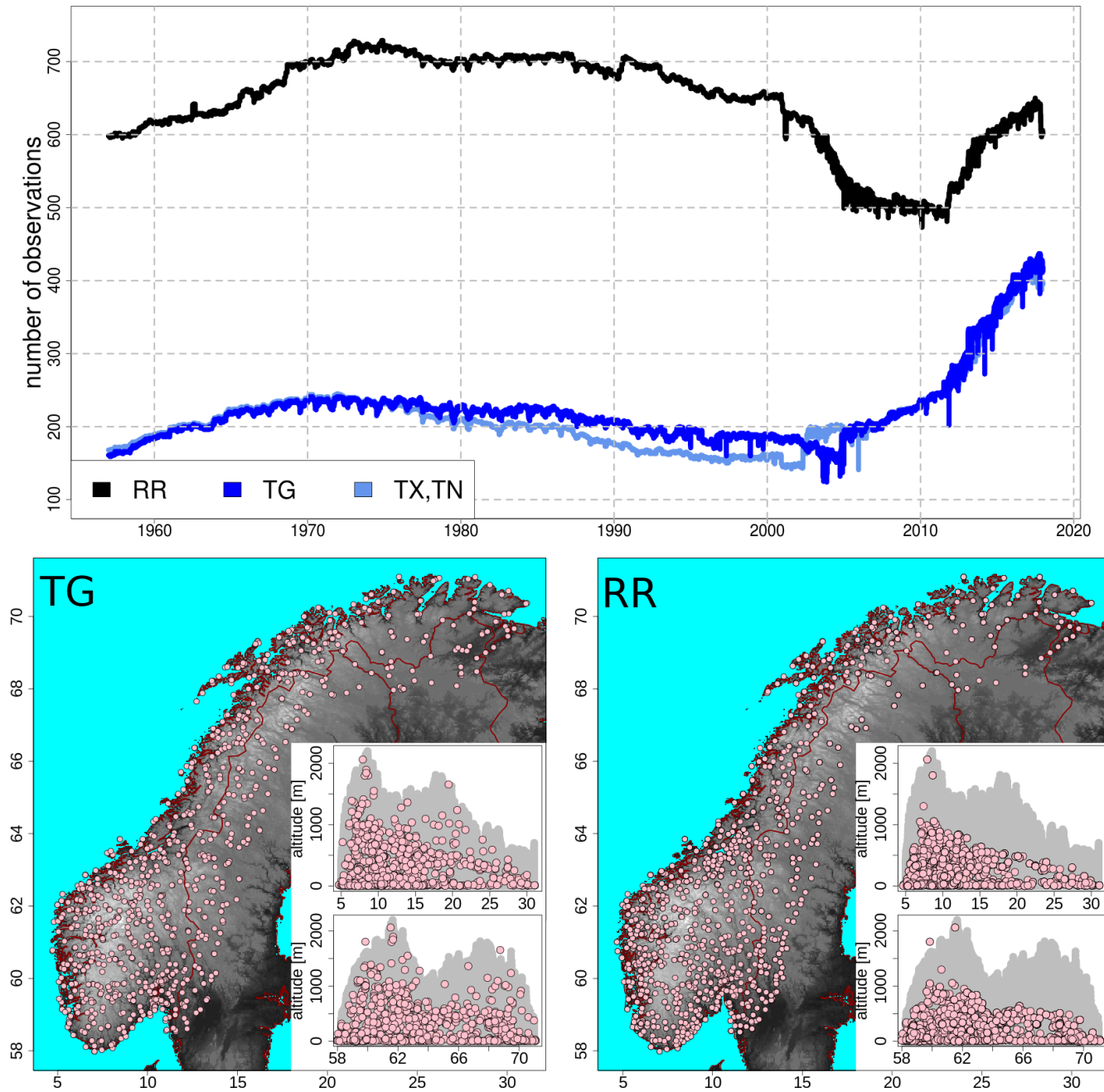
Figure 1. The observational network for the four variables: RR, TG, TX and TN. ~~A detailed description is given in Sec. ??.~~ The top panel shows the time series for the number of available observations over the Norwegian mainland. On the bottom panels, TG (left) and RR (right) observational networks are shown (TX and TN are similar to TG), with the pink dots marking station locations with more than 1 year of available data. The ~~displayed fields are the IDI at gridpoints~~geographic coordinate system is used. ~~Panels *a* (TG) and *b*~~ Both panels show the same geographical map of the domain with topographic information derived from a high-resolution digital elevation model (~~RR~~DEM)~~show~~. For each of the ~~CV-IDI~~two bottom panels, the two inset graphs show altitude above mean sea level as a function of ~~the distance to the nearest station~~latitude (top graph) and longitude (bottom graph). In ~~panel *c*~~each graph, the ~~two parameters measuring~~ gray area shows the ~~local station influence on~~altitude of the ~~analysis, CV-IDI and IDI~~terrain at gridpoints, while the pink dots are ~~compared~~the altitudes of stations.

22
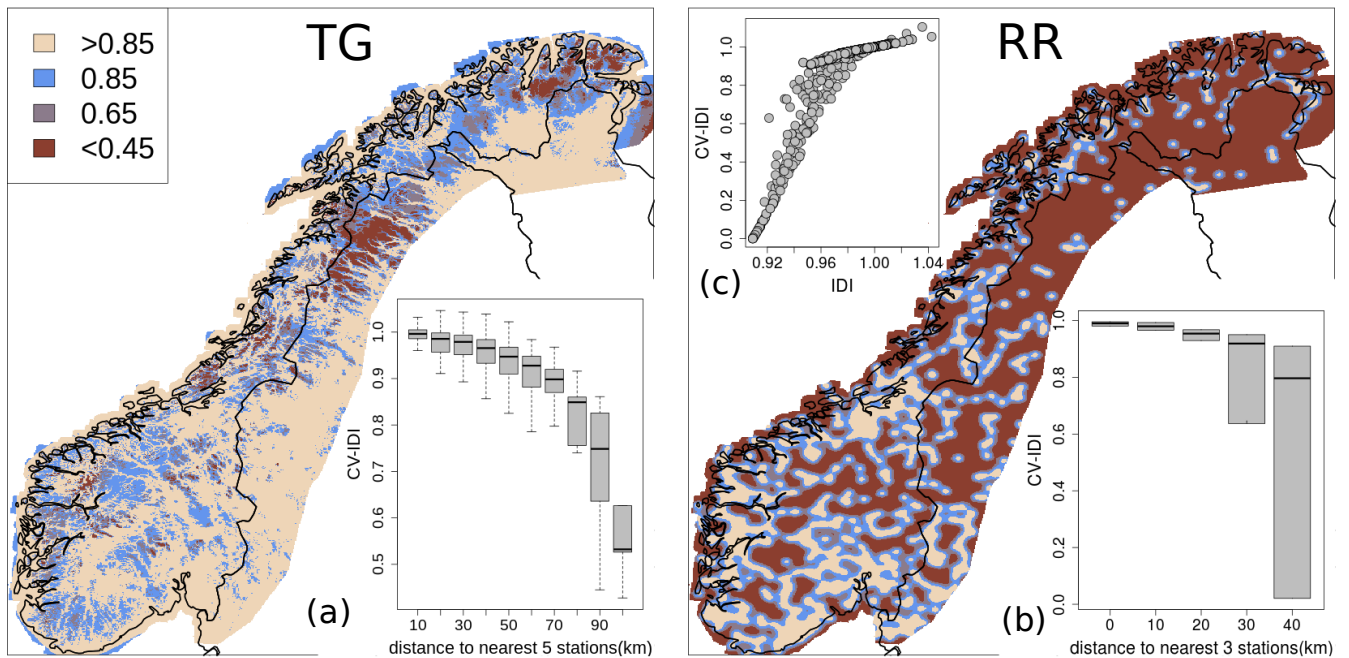
Figure 2. A representation of the observational network useful for spatial analysis of the variables: TG (left) and RR (right). As for Fig. **??**, TX and TN are similar to TG. The displayed fields are the Integral Data Influence (IDI) at gridpoints (same color scale for both fields). Panels *a* (TG, inset) and *b* (RR, inset top) show the Cross-Validation IDI (CV-IDI) as a function of the distance to the nearest stations. In panel *c* (RR, inset bottom), CV-IDI and IDI are compared.
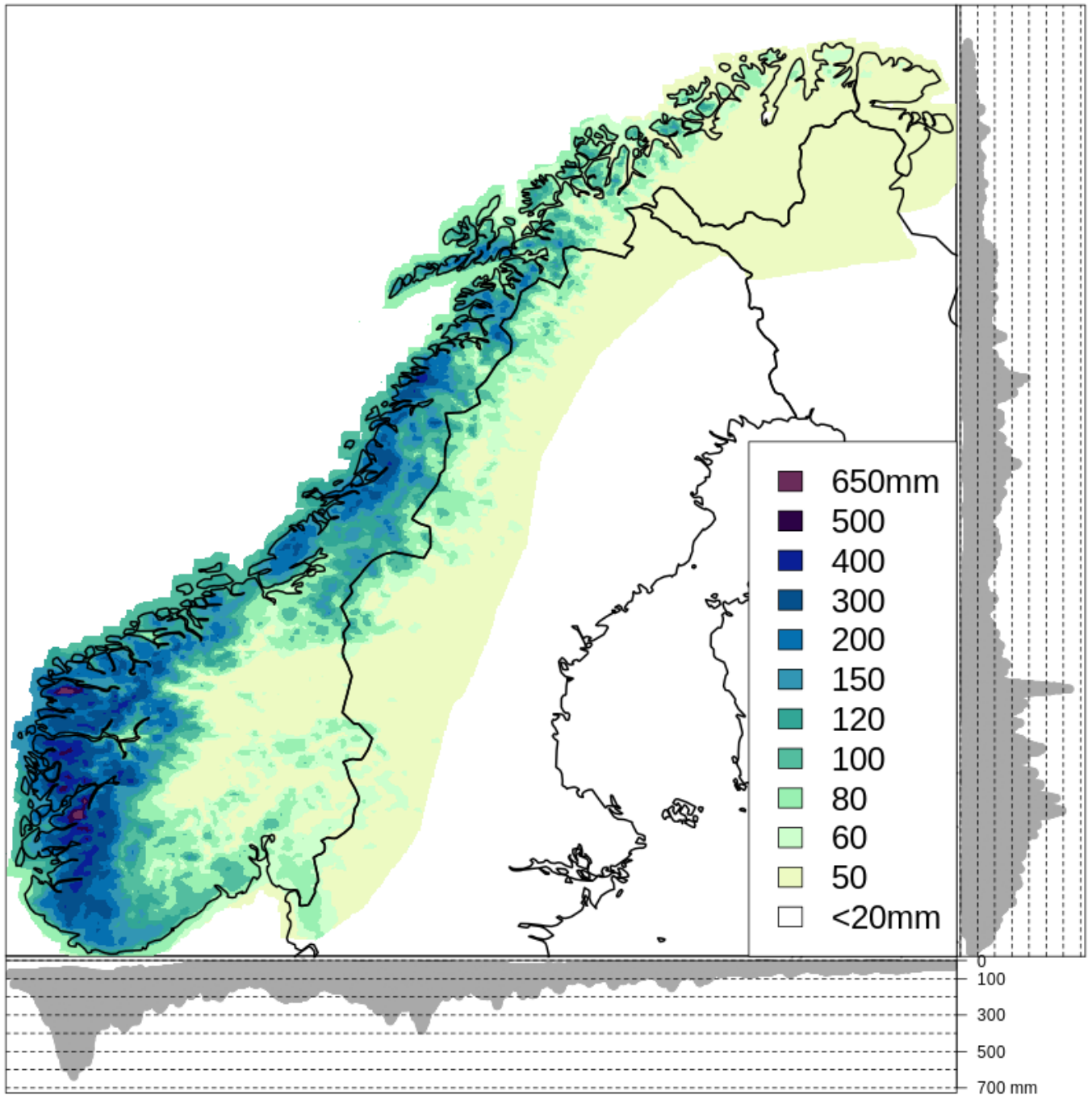
**Figure 3.** Precipitation reference field for October, that is x^{ref} in Eq. (**??**) used for 1998-10-24. The lateral and bottom panels in both graphs show the projection of the differences on the y- and the x- axis, respectively.

24
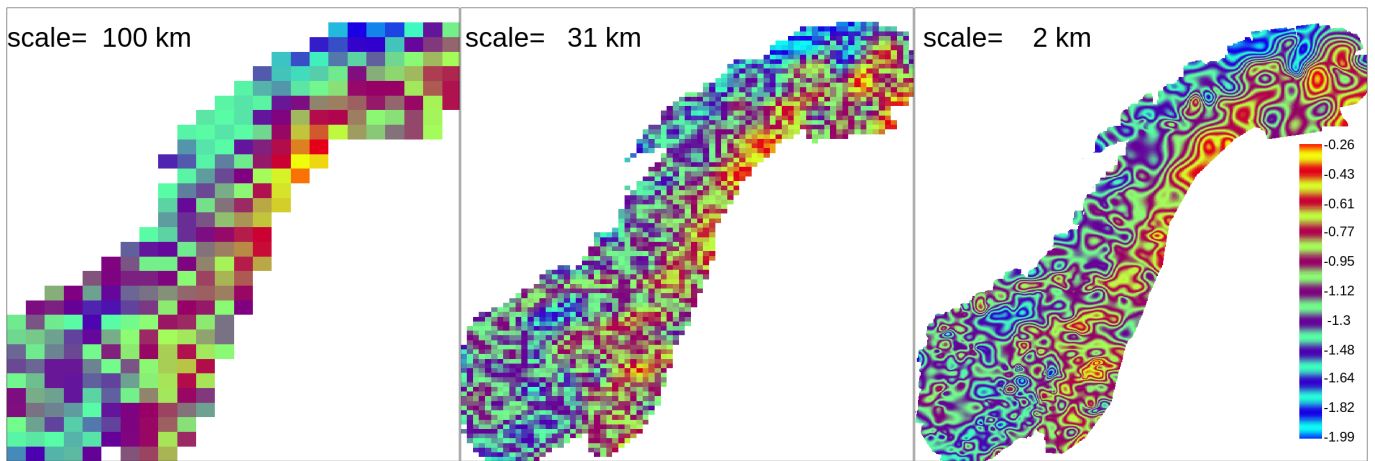
Figure 4. 1998-10-24 RR. Statistical interpolation of relative anomalies over different spatial scales. The statistical interpolation loops over 91 scales: 100 km is the seventeenth; 31 km is the sixty-ninth; 2 km is the ninety-first. The colour scale highlights the details in the field and it is the same for the three maps.
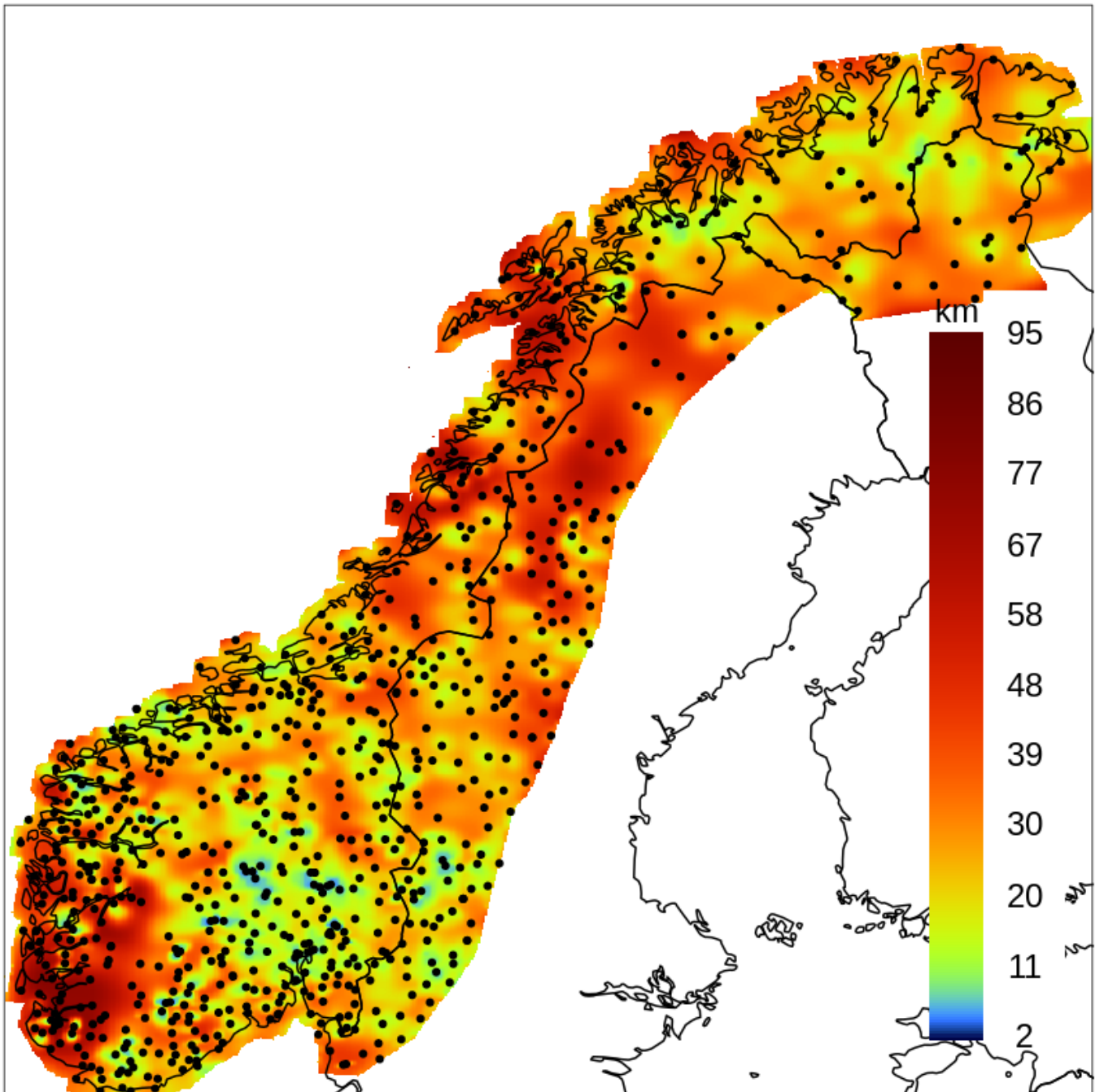
Figure 5. 1998-10-24 RR. With respect to Fig. ??, for each gridpoint the field shows the spatial scale where the last significant variation of the interpolated precipitation relative anomaly has occurred. The black dots mark the station locations.
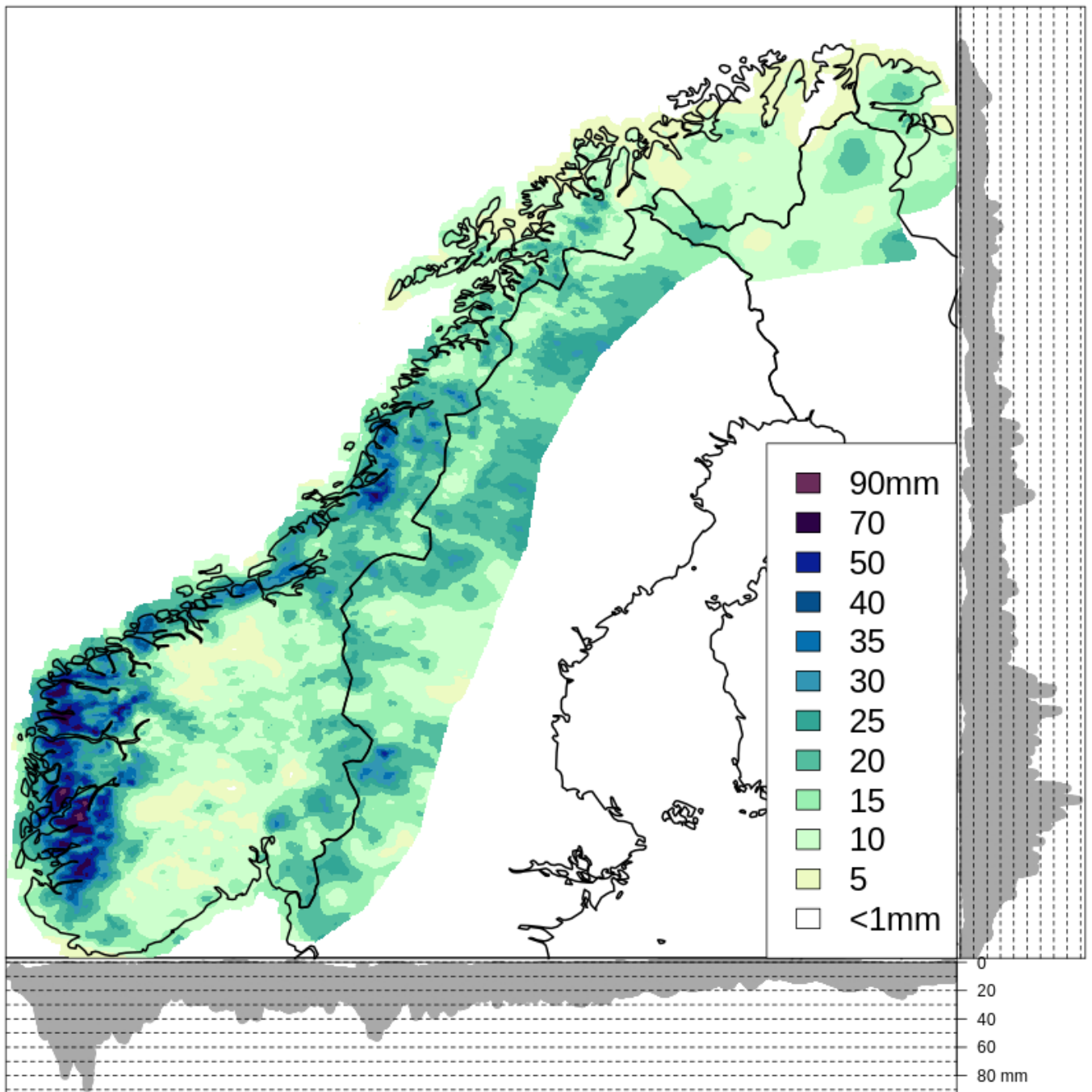
Figure 6. 1998-10-24, RR analysis field. The lateral and bottom panels in both graphs show the projection of the differences on the y- and the x- axis, respectively.
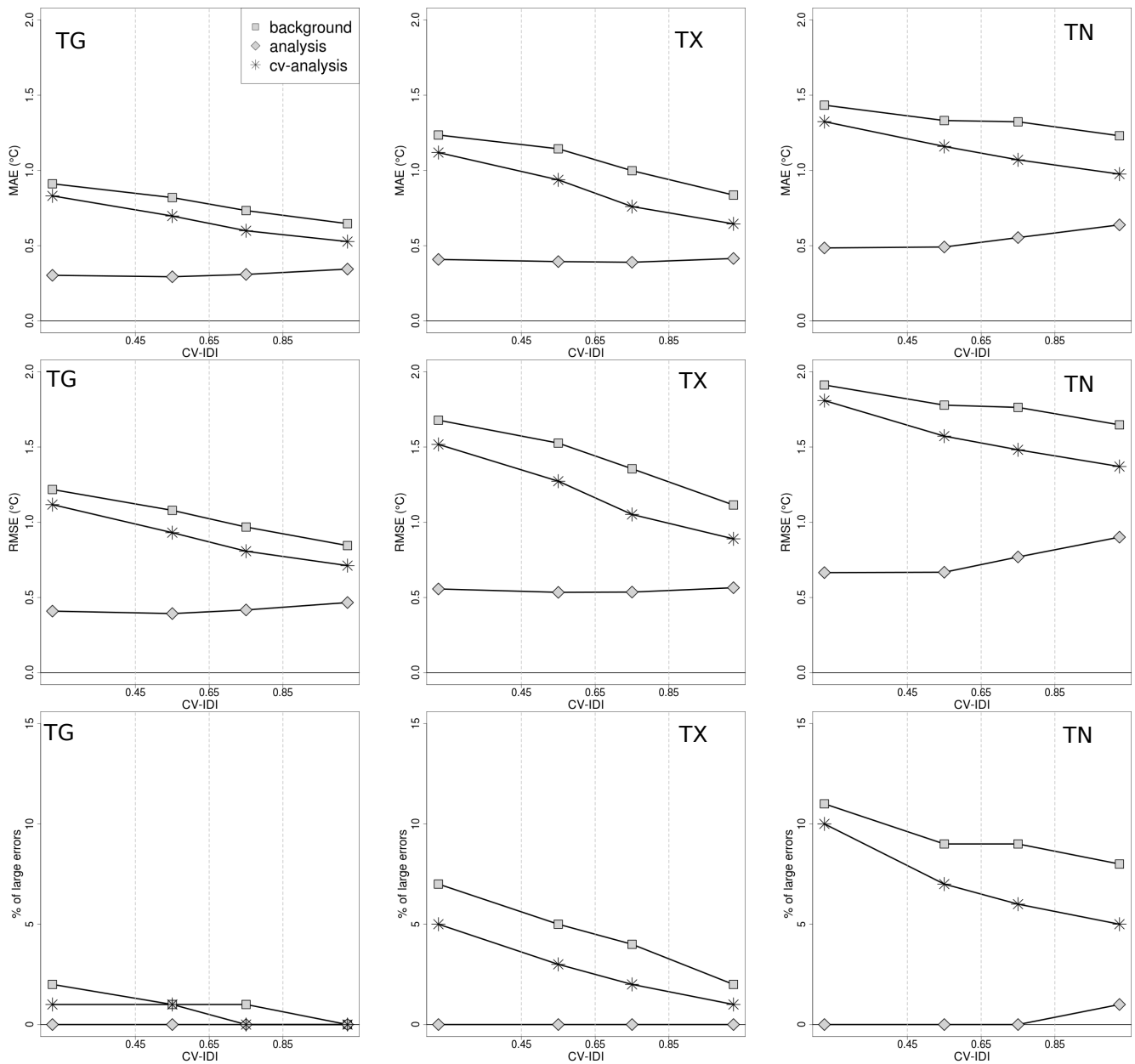
Figure 7. TG, TX and TN, verification scores as a function of CV-IDI based on the summer seasons (June-July-August) within the 61-year time period 1957-2017. The ~~verification~~ scores~~are based~~: for the analysis are based on the analysis residuals; for the CV-analysis on the CV-analysis residuals; for the background on the innovation. On the top row, the mean absolute error (MAE). In the middle, the root mean square error (RMSE). On the bottom row, the percentage of large errors. A large error is defined as the absolute value of innovation or residual larger than $3°$C.
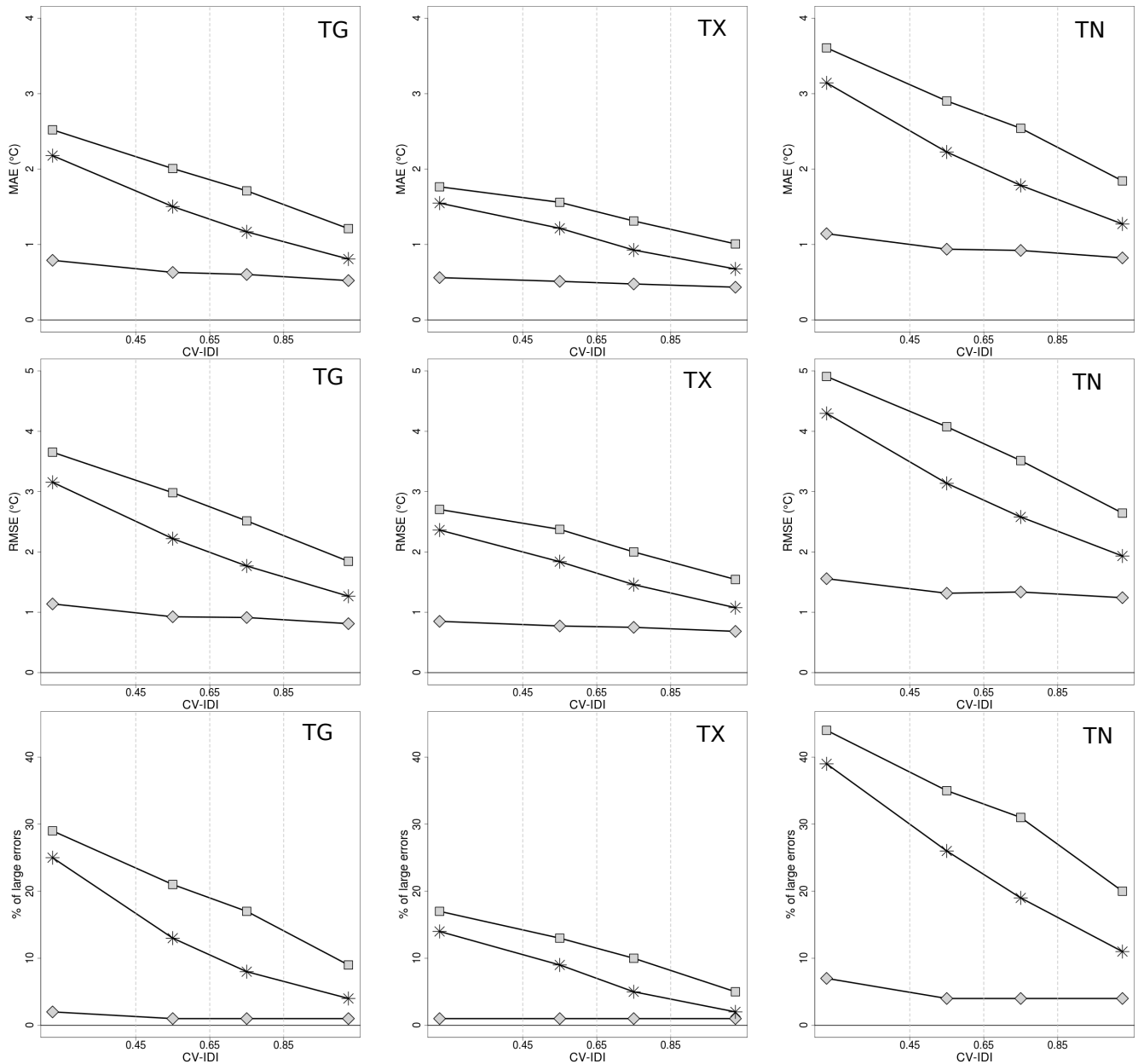
Figure 8. TG, TX and TN, verification scores as a function of CV-IDI based on winter seasons (December-January-February) within the 61-year time period 1957-2017. See Fig. **??**.
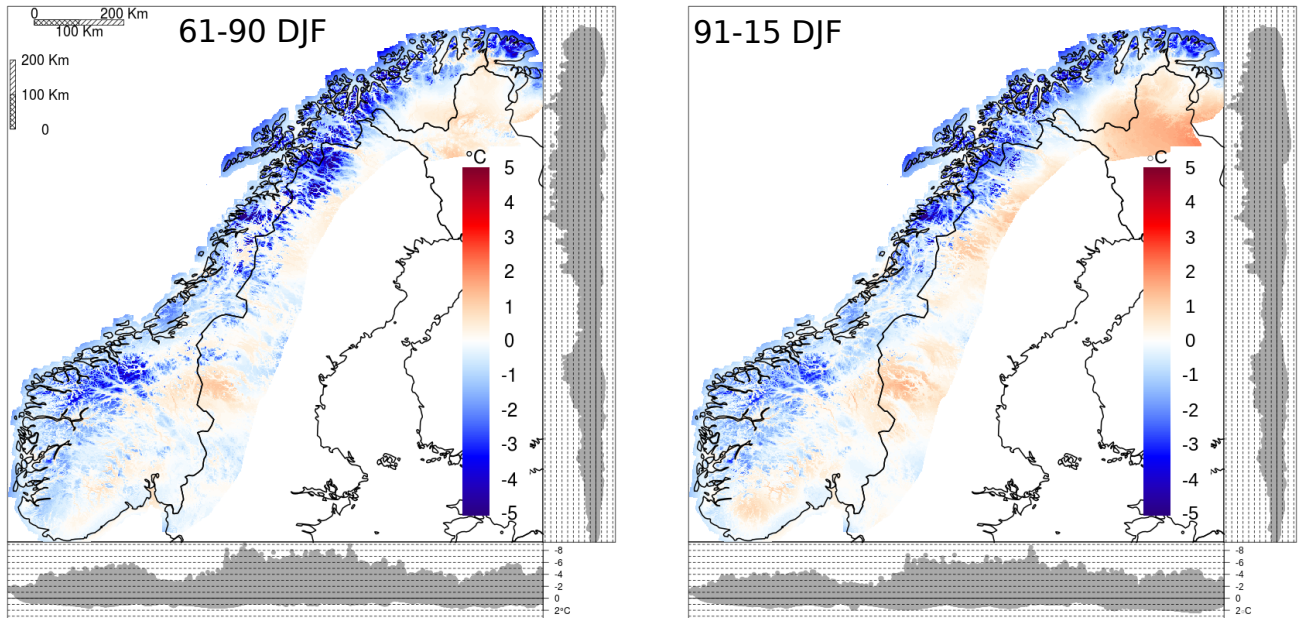
Figure 9. mean TG difference between seNorge_2018 and seNorge2 based on daily analysis in December, January and February. On the left panel, the 30-year period 1961-1990 is considered. On the right, the mean is based on the 25-year period 1991-2015. The lateral and bottom panels in both graphs show the projection of the differences on the y- and the x- axis, respectively.
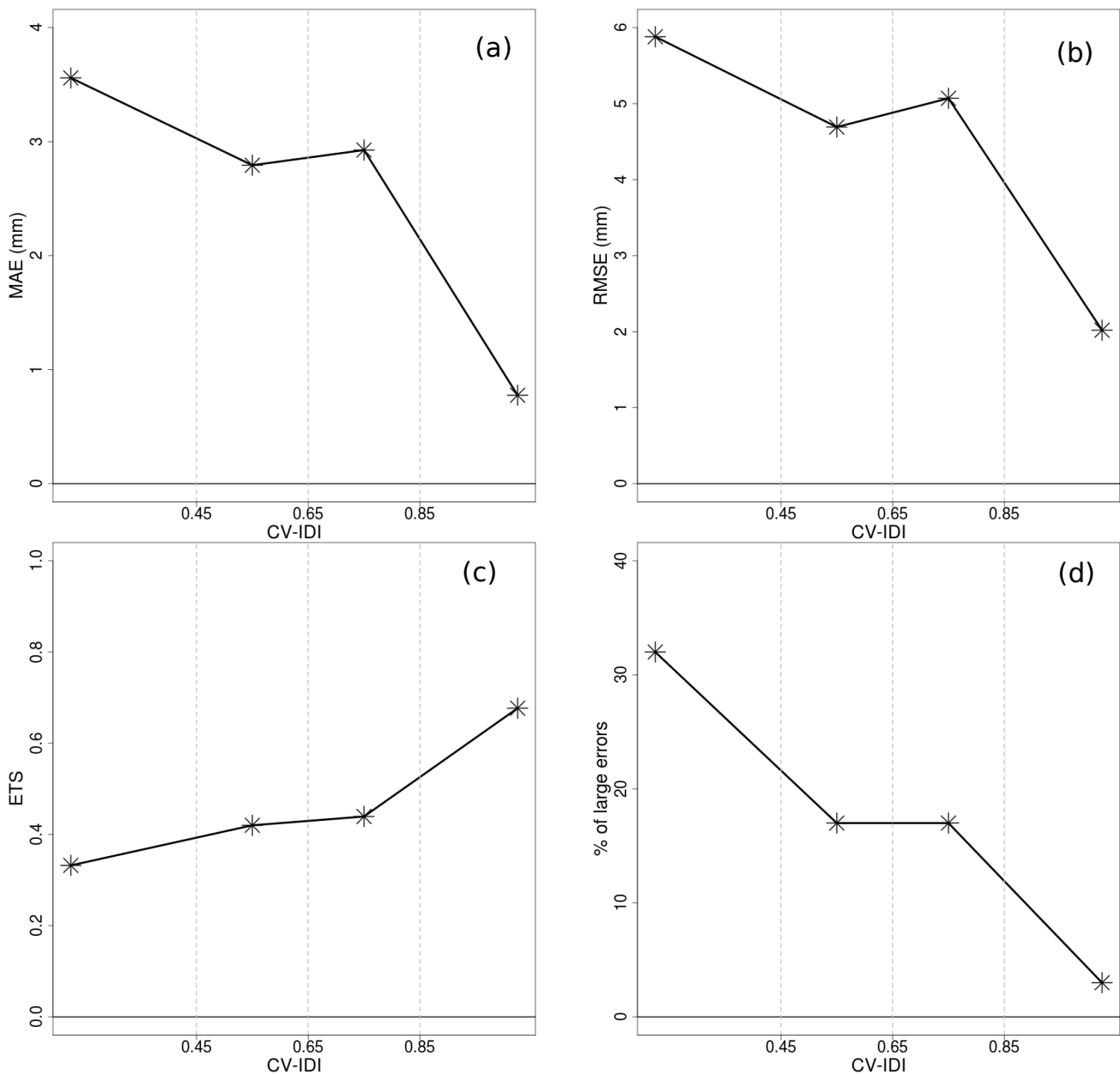
Figure 10. RR, CV-analysis scores as a function of CV-IDI based on CV-analysis residuals in the 61-year time period 1957-2017. Panel a, Mean Absolute Error considering only observations greater than 1 mm/day. Panel b, Root-Mean-Squared Error considering only observations greater than 1 mm/day. Panel c, Equitable Threat Score with threshold equals to 1 mm/day. Panel d, Percentage of large errors in case of intense precipitation (i.e., greater than 10 mm/day). A large error is defined as the absolute value of a CV-analysis residual larger than 50% of the observed value.
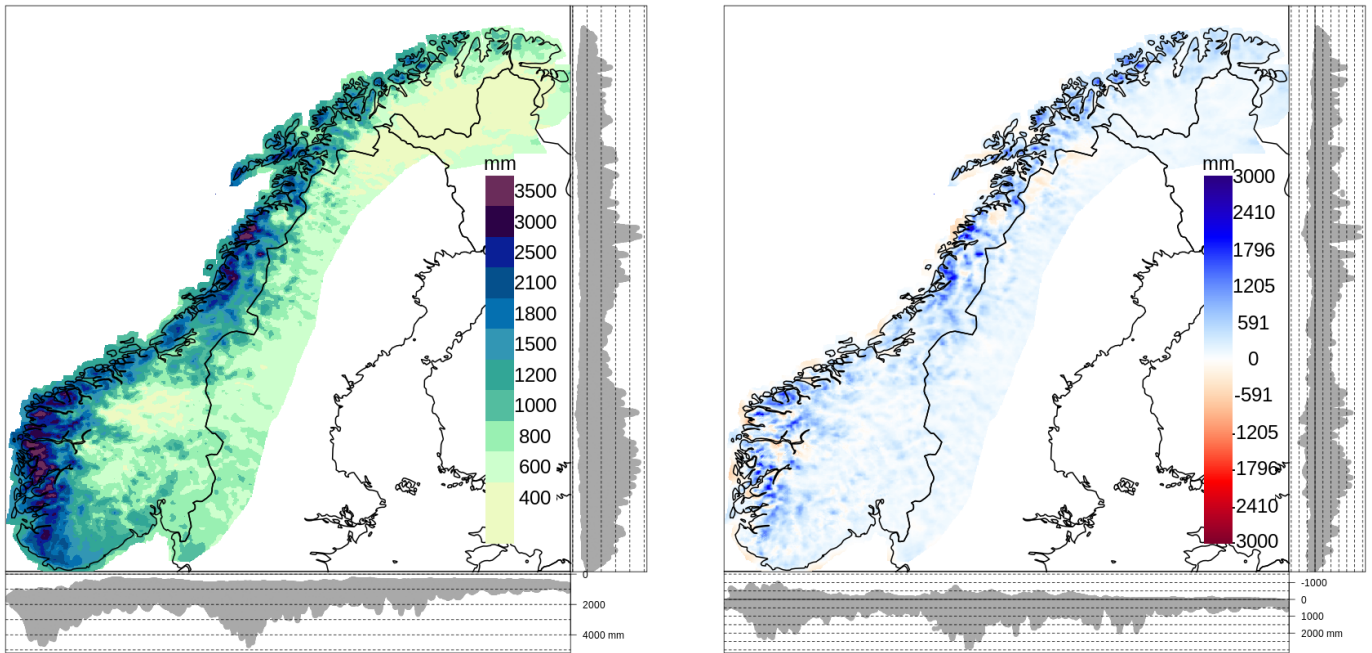
Figure 11. RR annual total precipitation. On the left, the mean annual total precipitation based on the 61-year period 1957-2017: the lower precipitation class includes values smaller than 400 mm; the upper precipitation class values between 3500 mm and 4700 mm. On the right, mean annual total precipitation difference between seNorge_2018 and seNorge2 based on the 51-year period 1957-2015. On each graph, the lateral and bottom panels show the projection of the differences on the y- and the x- axis, respectively.
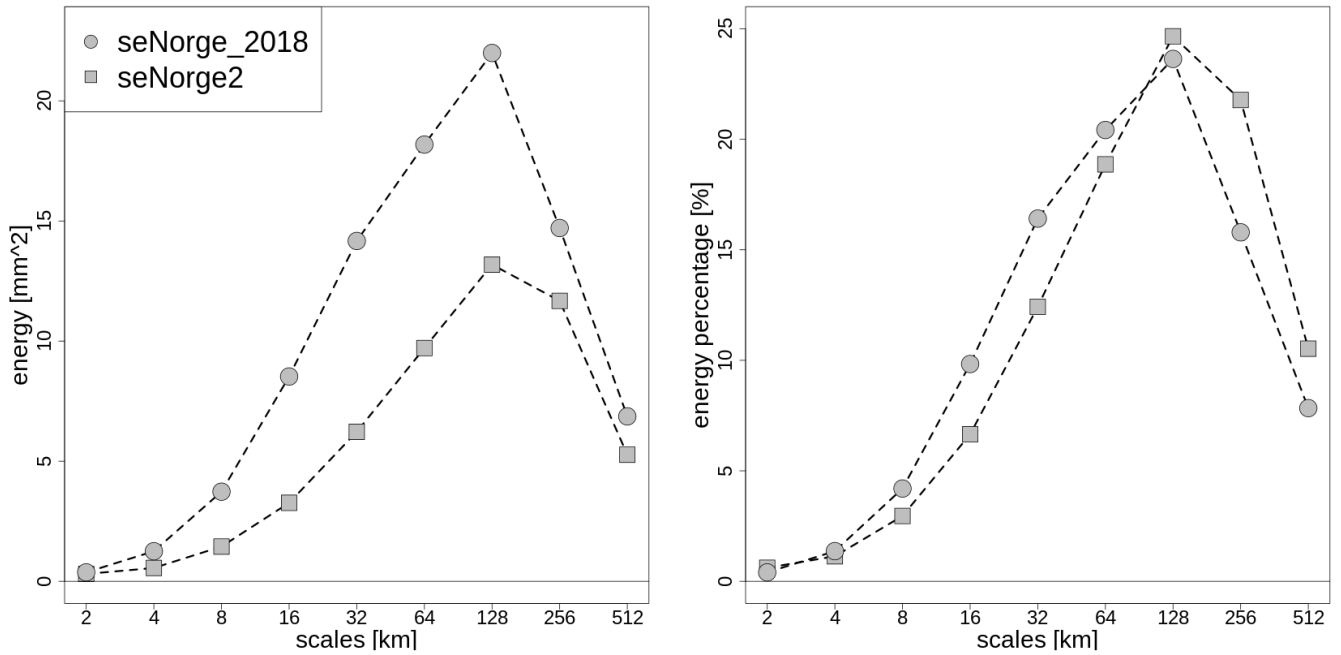
Figure 12. Scale decomposition of precipitation energy based on 2D discrete Haar wavelet transformation. On the left, the averaged energy as a function of the spatial scale for seNorge_2018 and seNorge2. On the right, the averaged percentage of squared energy as a function of the spatial scale for the same datasets. The statistics is based on the 25% of cases with the most intense average precipitation over the domain.
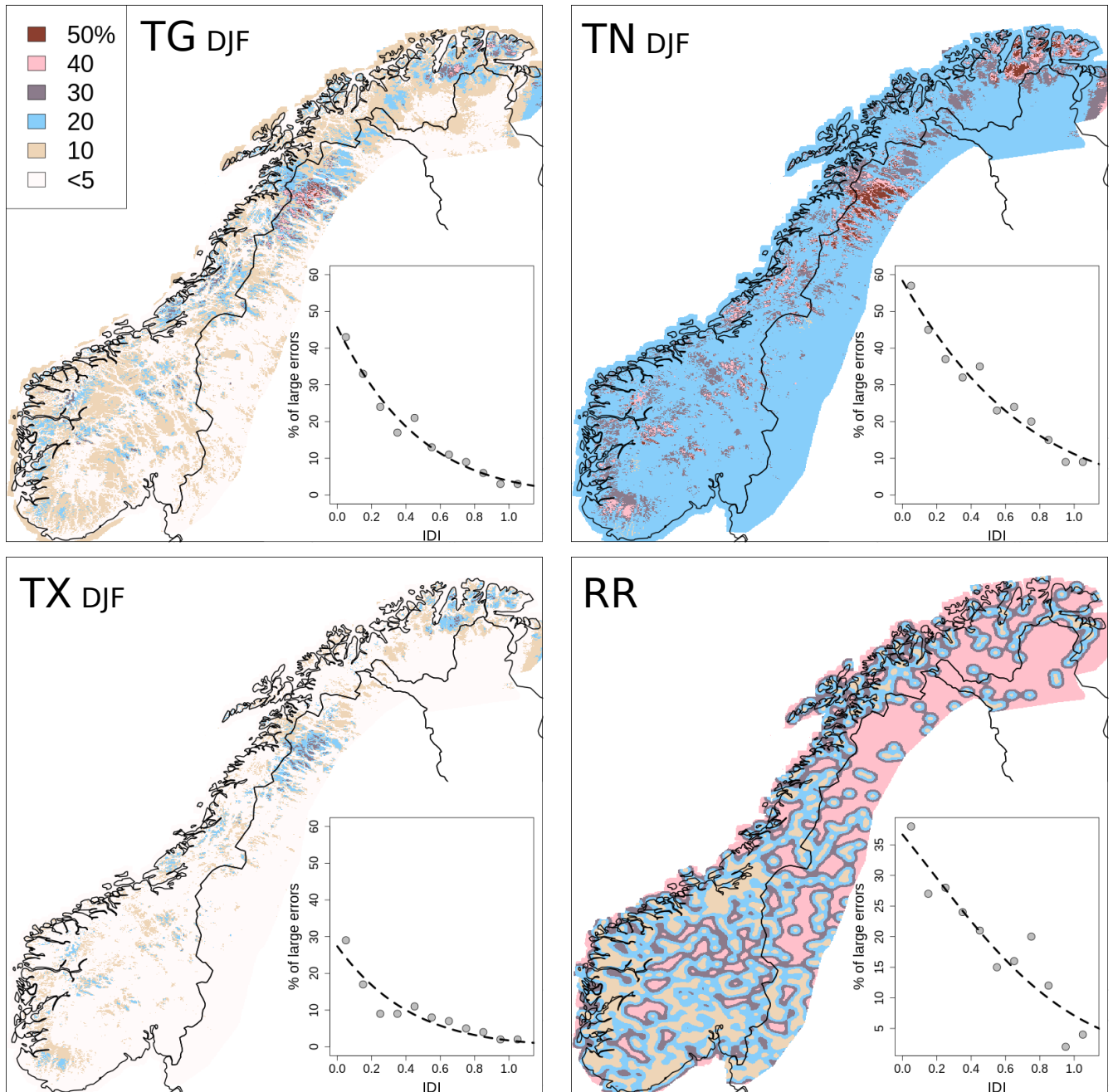
Figure 13. Expected percentage of large errors on the grid. The colour scale is the same for all the maps. Wintertime (DJF) temperatures are considered, large errors are defined as deviations between analysis and unknown truth larger than $3°$C. All precipitation data has been considered, large errors are deviations between analysis and unknown truth larger than 50% when the analysis value is greater than 10 mm/day. The insets show the relation between IDI and percentage of large errors: (the dots)correspond to the percentage of large errors observed at station locations (on the x-axis, CV-IDI instead of IDI); (line) the dashed lines are the best-fit function used to infer the expected percentage of large errors at gridpoints.

34

Table 1. TG annual statistics: "n" is the average number of stations; "d" (km) is the average distance between a station and its nearest third station; D^h (km), D^z (m), σ_b^2 $((^\circ C)^2)$, and σ_o^2 $((^\circ C)^2)$ are the optimal values given the 1-year innovation statistics and the constraint $\sigma_o^2/\sigma_b^2 = 0.5$.

year	n	d	D^h	D^z	σ_b^2	σ_o^2
1960	398	55	60	206	2.24	1.12
1970	669	42	45	217	1.86	0.93
1980	639	42	58	201	2.45	1.22
1990	600	44	57	202	1.33	0.66
2000	627	44	55	206	1.28	0.64
2010	639	45	52	206	2.45	1.23

Table 2. TX annual statistics (see Table. **??**).

year	n	d	D^h	D^z	σ_b^2	σ_o^2
1960	395	55	56	207	2.09	1.05
1970	669	42	57	201	1.67	0.84
1980	616	45	37	216	2.09	1.05
1990	563	47	55	206	1.40	0.70
2000	596	46	56	210	1.32	0.66
2010	638	45	57	206	2.09	1.04

Table 3. TN annual statistics (see Table. **??**).

year	n	d	D^h	D^z	σ_b^2	σ_o^2
1960	396	55	50	217	4.42	2.21
1970	670	42	53	222	3.80	1.90
1980	615	45	62	211	4.58	2.29
1990	560	47	52	210	2.88	1.44
2000	596	46	51	210	2.99	1.49
2010	637	45	64	212	4.70	2.35