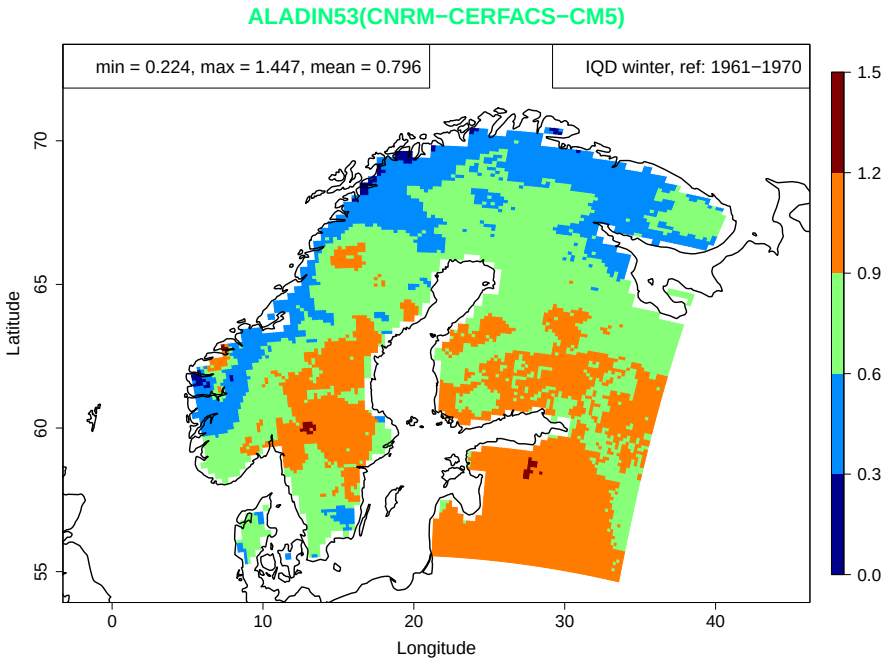
Norsk Regnesentral
NORWEGIAN COMPUTING CENTER

Report

Influence of reference period on evaluation of temperature output from the EURO-CORDEX climate ensemble using E-OBS data



Report no 1046

Authors Marion Haugen
Thordis L. Thorarinsdottir

Date 28th January 2020

ISBN 82-539-0556-3

Norsk Regnesentral
NORWEGIAN COMPUTING CENTER

The authors

Marion Haugen is Senior Research Scientist and Thordis L. Thorarinsdottir is Chief Research Scientist at Norwegian Computing Center.

Norwegian Computing Center

Norsk Regnesentral (Norwegian Computing Center, NR) is a private, independent, non-profit foundation established in 1952. NR carries out contract research and development projects in information and communication technology and applied statistical-mathematical modelling. The clients include a broad range of industrial, commercial and public service organisations in the national as well as the international market. Our scientific and technical capabilities are further developed in co-operation with The Research Council of Norway and key customers. The results of our projects may take the form of reports, software, prototypes, and short courses. A proof of the confidence and appreciation our clients have in us is given by the fact that most of our new contracts are signed with previous customers.

Norsk Regnesentral
Norwegian Computing Center
Postboks 114, Blindern
NO-0314 Oslo, Norway

Besøksadresse
Office address
Gaustadalléen 23
NO-0373 Oslo, Norway

Telefon · telephone
(+47) 22 85 25 00
Telefaks · telefax
(+47) 22 69 76 60

Internett · internet
www.nr.no
E-post · e-mail
nr@nr.no

Title	**Influence of reference period on evaluation of temperature output from the EURO-CORDEX climate ensemble using E-OBS data**
Authors	**Marion Haugen** <marionh@nr.no> **Thordis L. Thorarinsdottir** <thordis@nr.no>
Date	28th January 2020
ISBN	82-539-0556-3
Publication number	1046

Abstract

This note investigates the fit of daily temperature projections over Fennoscandia for 1950-2005 for nine combinations of global and regional climate models from EURO-CORDEX. The evaluation is performed by comparing temperature anomalies from the climate models to observed temperature anomalies from the E-OBS dataset using integrated quadratic distance (IQD). Summer and winter are evaluated separately. Different alternatives for a reference period are explored; 30-year periods, 20-year periods and 10-year periods.

The evaluation results are robust for the 30-year periods, in both seasons. When the length of the reference period decreases, the evaluation results become less and less robust. It is hard to say if 20 years are sufficient so it is better to use 30 years. A reference period of 10 years is too short because the evaluation results highly depend on the chosen period.

Keywords	Climate model evaluation, integrated quadratic distance, reference period, temperature anomalies
Target group	Climate scientists, users of climate information
Availability	Open
Project	PostClim
Project number	220778
Research field	Statistics, climate science
Number of pages	29

Contents

1 Introduction

Meteorological measured parameters such as temperature or precipitation vary greatly month to month and year to year. To describe the climate of a location, these measurements are averaged over a prolonged period of time. The calculated averages are referred to as normal values or climate normals, and comprise the reference (or baseline) period for the evaluation of anomalies in climate variability and climate change monitoring.

A temperature anomaly is the difference from a long-term average, or baseline, temperature. The baseline temperature is typically computed by averaging 30 years of temperature data. Standard 30-year reference periods, according to the World Meteorological Organization (WMO), are 1901-1930, 1931-1960 and 1961-1990. A positive anomaly indicates the observed temperature was warmer than the baseline, while a negative anomaly indicates the observed temperature was cooler than the baseline.

In view of the progressive climate change in recent decades, the question arises about the representativeness of a reference period such as 1961-1990. It has been argued that the reference period should be a set of 30 years updated every ten years (Wright, 2012), e.g. 1971-2000 and 1981-2010. The purpose of this is to provide climate normals that adequately describe the current climate and these may also be used as expected values.

In this note we will evaluate daily temperature projections over Fennoscandia for the years from 1950 to 2005 for nine combinations of global and regional climate models[1] from EURO-CORDEX. The integrated quadratic distance (IQD) will be used for evaluating the temperature anomalies from the climate models against the temperature anomalies from the E-OBS dataset. This will be performed separately for the summer and winter seasons. We explore different alternatives for a reference period and answer the following questions:

- If we use a set of 30 years, does it matter whether we use 1951-1980, 1961-1990 or 1971-2000?
- What happens if we use 20 years instead of 30?
- What if we only use 10 years?

The remainder of the note is organized as follows. Section 2 describes the applied datasets. Section 3 explains the theory behind IQD. Section 4 explains what was done to obtain our results, which are presented in Section 5. Finally, the conclusion is provided in Section 6. Plots displaying model rankings for selected reference periods are given in Appendix A. Plots displaying the mean IQD over Fennoscandia are given in Appendix B for the summer season and Appendix C for the winter season.

1. Global climate models (GCM) are used for projecting climate changes on a global scale. To project local climate changes with some precision, regional climate models (RCM) with finer resolution and boundary conditions from a GCM have to be constructed.

2 Data

2.1 Climate models

A total of nine different combinations of global and regional climate models from EURO-CORDEX are chosen for the evaluation. The models are given in Table 1. Each model contains daily temperature data measured at the Kelvin scale from various start and end dates. All climate models follow the standard Gregorian calendar except models 8 and 9 which have 30 days in all months.

Table 1. Nine GCM/RCM combinations from EURO-CORDEX used in our testing.

Model nr.	Global climate model	Ensemble member	Regional climate model	Institute	Institution name
1	CNRM-CERFACS-CM5	r1i1p1	CCLM4-8-17	CLMcom	Climate Limited-area Modelling Community
2	CNRM-CERFACS-CM5	r1i1p1	ALADIN53	CNRM	Météo-France / Centre National de Recherches Météorologiques
3	ICHEC-EC-EARTH	r1i1p1	RACMO22E	KNMI	Royal Netherlands Meteorological Institute
4	ICHEC-EC-EARTH	r12i1p1	CCLM4-8-17	CLMcom	Climate Limited-area Modelling Community
5	MPI-ESM-LR	r1i1p1	CCLM4-8-17	CLMcom	Climate Limited-area Modelling Community
6	MPI-ESM-LR	r1i1p1	REMO2009	MPI-CSC	Helmholtz-Zentrum Geesthacht, Climate Service Center, Max Planck Institute for Meteorology
7	MPI-ESM-LR	r2i1p1	REMO2009	MPI-CSC	Helmholtz-Zentrum Geesthacht, Climate Service Center, Max Planck Institute for Meteorology
8	MOHC-HadGEM2-ES	r1i1p1	CCLM4-8-17	CLMcom	Climate Limited-area Modelling Community
9	MOHC-HadGEM2-ES	r1i1p1	RACMO22E	KNMI	Royal Netherlands Meteorological Institute

The climate models have data for the latitude/longitude grid shown in Figure 1. We will analyze Fennoscandia, which is the geographical peninsula of the Nordic region comprising the Scandinavian Peninsula, Finland, Karelia, and the Kola Peninsula.

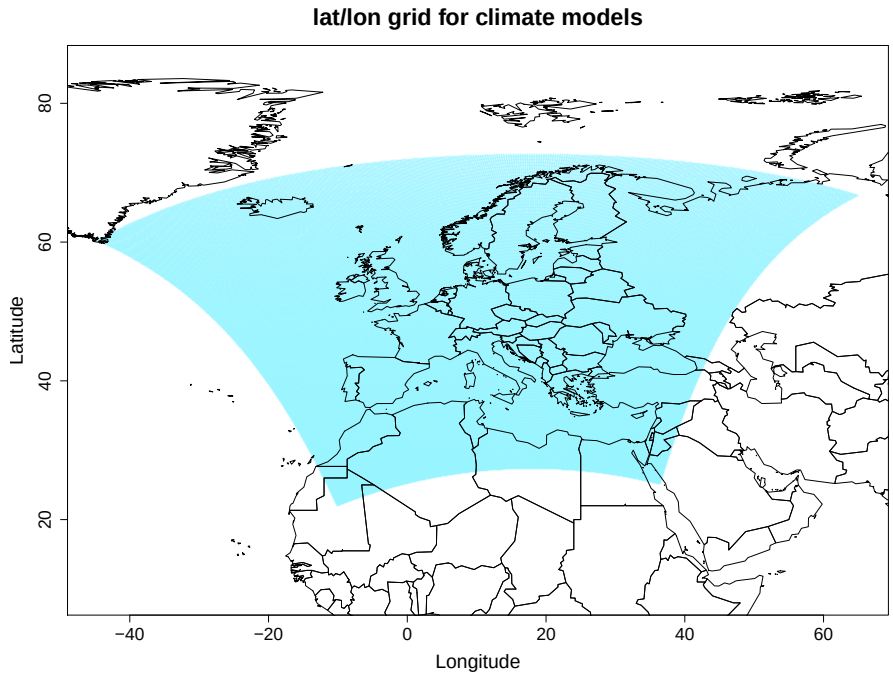


Figure 1. Grid with latitude and longitude values for the EURO-CORDEX climate models.

2.2 Observations

The observed temperature data is collected from the E-OBS dataset (Haylock et al., 2008), a daily gridded observational dataset for precipitation, temperature and sea level pressure in Europe for 1950-2006. The daily mean temperature is measured at the Celsius scale, follows the Gregorian calendar and covers the latitude/longitude grid shown in Figure 1.

2.3 Preparation of data

All data files are in compressed NetCDF format and have to be prepared before the analysis, both for the climate models and E-OBS. Several files are merged to retrieve data for the entire period 1950-2005. Data before 1950, after 2005 and February 29 is removed. February 29 is not removed from the climate models with 30 days in each month (models 8 and 9). All the data from the climate models is converted from Kelvin to Celsius. Additionally, we are only interested in the 140x155 grid covering Fennoscandia. To be able to separate between land and sea in the climate models, NA values are inserted from the E-OBS data wherever there is water in the grid we examine.

For each individual time series, daily mean values for a given reference period (see Section 4.1) are calculated and the climate normals are subtracted from the original data, giving temperature anomalies. Anomalies for seasons are then extracted. Summer consists of June, July and August (abbreviated JJA). Winter consists of December, January and February (abbreviated DJF). This gives a total of 92 observations per summer and 90 observations per winter. The climate models with 30 days in each month have 90 observations for both seasons.

3 Theory

3.1 IQD

We denote a temperature observation by $y \in \Omega$ where Ω denotes the real axis $\mathbb{R}$. A probabilistic prediction for y is given by a distribution function with support on Ω denoted by $F \in \mathcal{F}$ for some appropriate class of distributions $\mathcal{F}$, with the density denoted by f if it exists.

Scoring rules assess the accuracy of probabilistic predictions by assigning a numerical penalty to each prediction-observation pair. Specifically, a scoring rule is a mapping

$$S : \mathcal{F} \times \Omega \to \mathbb{R} \cup \{\infty\} \tag{1}$$

where, in our notation, a smaller penalty indicates a better prediction. A scoring rule is *proper* relative to the class $\mathcal{F}$ if

$$\mathbb{E}_G S(G, Y) \leq \mathbb{E}_G S(F, Y) \tag{2}$$

for all probability distributions $F, G \in \mathcal{F}$, that is, if the expected score for a random observation Y is optimized if the true distribution of Y is issued as the prediction. The scoring rule is *strictly proper* relative to the class $\mathcal{F}$ if (2) holds with equality only if $F = G$. Propriety will encourage honesty and prevent hedging, which coincides with Murphy's first type of goodness (Murphy, 1993).

In some cases, in particular in climate modelling, it is of interest to compare the predictive distribution F against the true distribution of the observations which is commonly approximated by the *empirical distribution function* of the available observations $y_1, \ldots, y_n$,

$$\widehat{G}_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{y_i \leq x\}. \tag{3}$$

The two distributions, F and $\widehat{G}_n$, can be compared using a *divergence*

$$D : \mathcal{F} \times \mathcal{F} \to \mathbb{R} \tag{4}$$

where $D(F, F) = 0$.

Assume that the observations $y_1, \ldots, y_n$ forming the empirical distribution function $\widehat{G}_n$ are independent with distribution $G \in \mathcal{F}$. A propriety condition for divergences corresponding to that for scoring rules (2) states that the divergence D is *k-proper* for a positive integer k if

$$\mathbb{E}_G D(G, \widehat{G}_k) \leq \mathbb{E}_G D(F, \widehat{G}_k) \tag{5}$$

and *asymptotically proper* if

$$\lim_{k \to \infty} \mathbb{E}_G D(G, \widehat{G}_k) \leq \lim_{k \to \infty} \mathbb{E}_G D(F, \widehat{G}_k) \tag{6}$$

for all probability distributions $F, G \in \mathcal{F}$ (Thorarinsdottir et al., 2013). While the condition in (6) is fulfilled by a large class of divergences, only score divergences have been shown to fulfill (5) for all integers k. A divergence D is a *score divergence* if there exists a proper scoring rule S such that $D(F, G) = \mathbb{E}_G S(F, Y) - \mathbb{E}_G S(G, Y)$.

A score divergence that assesses the full distributions is the *integrated quadratic distance* (Thorarinsdottir et al., 2013):

$$\mathrm{IQD}(F, G) = \int_{-\infty}^{+\infty} \big(F(x) - G(x)\big)^2 \mathrm{d}x \tag{7}$$

In the following, we will apply the IQD to compare empirical distributions of climate model outputs and observations. A low IQD value is favorable and we expect the best (smallest) IQD if the model outputs and the data product values come from the same distribution. As a consequence, we can use IQD values to rank the performance of competing models.

4 Methods

4.1 Selected reference periods

Initially, the following 12 reference periods were tested:

- 30 years: 1951-1980, 1961-1990, 1971-2000 (3 in total)
- 20 years: 1951-1970, 1961-1980, 1971-1990, 1981-2000 (4 in total)
- 10 years: 1951-1960, 1961-1970, 1971-1980, 1981-1990, 1991-2000 (5 in total)

Temperature anomalies are calculated for each individual time series. E.g. for the reference period 1951-1980 anomalies for the climate models are differences between the temperature projection for climate model $1, 2, \ldots, 9$ and the long-term average temperature from 1951-1980 of the temperature projections for the same climate model. Anomalies for E-OBS are differences between the observed temperature and the 1951-1980 time-mean of the time series for E-OBS.

4.2 Rolling reference periods

Further, the following rolling reference periods were tested:

- 30 years: 1951-1980, 1952-1981, 1953-1982, . . ., 1971-2000 (21 in total, 18 new)
- 20 years: 1951-1970, 1952-1971, 1953-1972, . . ., 1981-2000 (31 in total, 27 new)
- 10 years: 1951-1960, 1952-1961, 1953-1962, . . ., 1991-2000 (41 in total, 36 new)

4.3 Evaluation

We want to evaluate the IQD for daily temperature anomalies for the E-OBS data and the nine climate models, separately for each model grid point. To do this we compare one year from E-OBS data with nine years of model data; four years before and four years after in addition to the year we want to examine. For the time period 1950-2005 this gives 48 comparisons per model, one for each year between 1954 and 2001:

- E-OBS data from 1954 compared to climate model data from 1950-1958
- E-OBS data from 1955 compared to climate model data from 1951-1959
- . . .
- E-OBS data from 2001 compared to climate model data from 1997-2005

For each grid point for Fennoscandia, we have 92 observations from E-OBS and nine times as many observations (828) from the climate models for each comparison when evaluating summer. For winter we have 90 and 810 observations, respectively. After calculating the IQD in all grid points for each of the 48 comparisons, we calculate the mean IQD in each grid point. The mean IQD over the whole grid for Fennoscandia is applied to make a ranking of the climate models, which is compared for the different reference periods.

5 Results

5.1 Model rankings

Mean IQD for daily summer temperature anomalies over all grid points for Fennoscandia, for 30-year, 20-year and 10-year rolling reference periods, is shown in Figure 2. Results for winter temperature anomalies are given in Figure 3. Figures A.1-A.3 show a scatter plot with ranking of the mean IQD for daily summer temperature anomalies for selected reference periods. Plots for winter temperature anomalies are given in Figures A.4-A.6. The nine climate models are ranked from lowest to highest overall mean IQD. All comparisons are relative to the reference period 1961-1990. If the rankings change, it is important to evaluate the possible change in the overall mean IQD value for the two reference periods, given on the right side of the y-axis in the scatter plots.

The evaluation results for summer and winter are similar. The rankings are stable over different 30-year reference periods, vary slightly for 20-year reference periods and are very unstable for 10-year reference periods. CCLM4-8-17(MOHC-HadGEM2-ES) performs best in the summer season and RACMO22E(MOHC-HadGEM2-ES) performs best in the winter season. ALADIN53(CNRM-CERFACS-CM5) performs worst in both seasons.

5.2 Model performance in grid points

To examine the spatial properties of the errors for the climate models with best and worst performance, the mean IQD at each grid point from these models is plotted on top of a map of Fennoscandia. To compare the results for the reference periods, all plots for one season have the same scale for the IQD value. These plots reveal the areas of Fennoscandia with higher IQD values, indicating a worse correspondence between modelled and observed temperature distributions. This is especially evident for the 10-year reference periods. Maps for six different reference periods (1961-1990, 1961-1980, 1971-1990, 1961-1970, 1971-1980 and 1981-1990) are given in Appendix B for the summer season and Appendix C for the winter season.

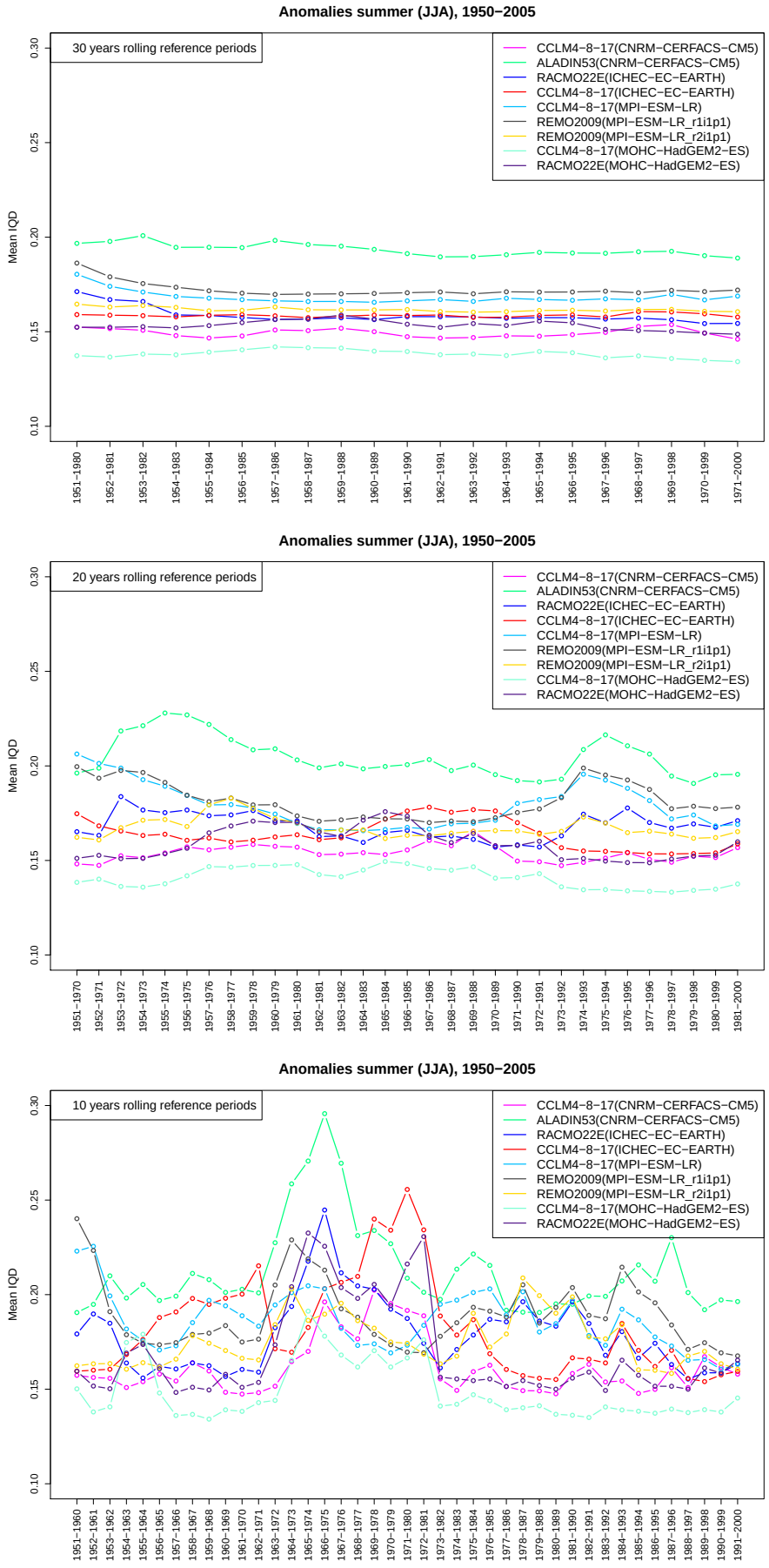


Figure 2. Mean IQD for daily summer temperature anomalies for nine climate models for rolling reference periods of 30, 20 and 10 years.

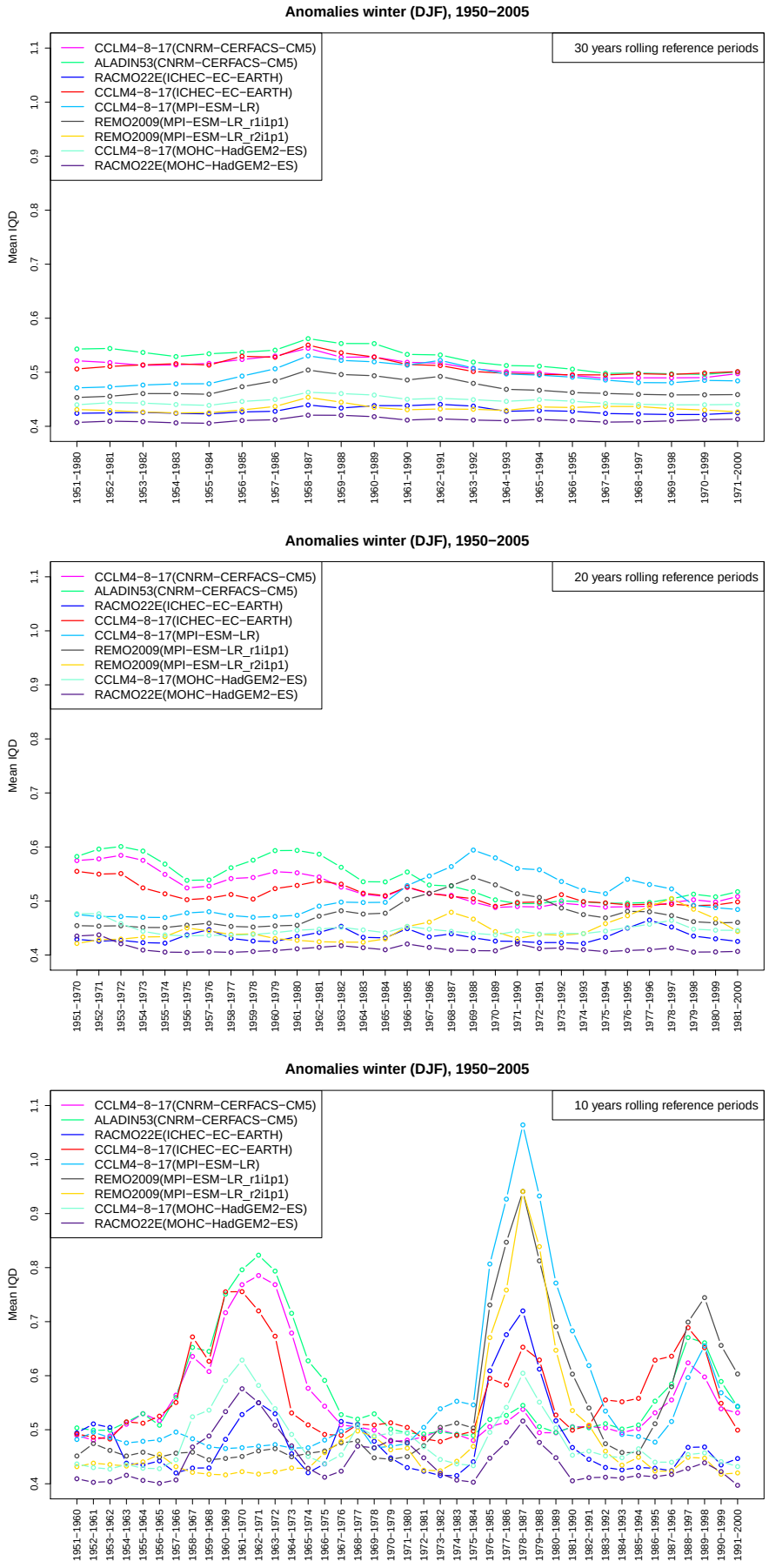


Figure 3. Mean IQD for daily winter temperature anomalies for nine climate models for rolling reference periods of 30, 20 and 10 years.

6 Conclusion

When climate model output is compared with observational data, it is standard practice to compare anomalies with respect to a reference period. In this note, we have demonstrated how the choice of reference period affects the ranking of climate models.

We have seen that different alternatives for reference period have a great influence on the results when evaluating the daily temperature projections over Fennoscandia for the years from 1950 to 2005. This applies both in summer and winter season. The evaluation results are robust for different 30-year reference periods but become less robust for 20-year reference periods. It is hard to say if 20 years are sufficient. We believe it is better to use 30 years. With 10-year reference periods, the evaluation results highly depend on the chosen period. We do not recommend to use 10 years.

Careful consideration of sensitivities to the choice of climate reference period is required to reliably compare climate models with observations and to produce robust projections of future climate (Hawkins and Sutton, 2016).

References

Hawkins, E. and Sutton, R. (2016). Connecting climate model projections of global temperature change with the real world. *Bulletin of the American Meteorological Society*, 97(6):963–980. 14

Haylock, M. R., Hofstra, N., Klein Tank, A. M. G., Klok, E. J., Jones, P. D., and New, M. (2008). A European daily high-resolution gridded data set of surface temperature and precipitation for 1950-2006. *Journal of Geophysical Research: Atmospheres*, 113:D20119. doi:10.1029/2008JD010201. 7

Murphy, A. H. (1993). What is a good forecast? An essay on the nature of goodness in weather forecasting. *Weather and Forecasting*, 8(2):281–293. 8

Thorarinsdottir, T. L., Gneiting, T., and Gissibl, N. (2013). Using proper divergence functions to evaluate climate models. *SIAM/ASA Journal on Uncertainty Quantification*, 1(1):522–534. 9

Wright, W. (2012). Discussion paper on the calculation of the standard climate normals: a proposal for a dual system. World Climate Data and Monitoring Programme. Available from: http://www.wmo.int/pages/prog/wcp/wcdmp/documents/Rev_discussion_paper_May2012.pdf. 5

A Model rankings with selected reference periods

Mean IQD for daily temperature anomalies for 11 different reference periods are compared to the mean IQD for daily temperature anomalies for the reference period 1961-1990 (given to the left in the figures). The nine climate models are ranked from lowest to highest overall mean IQD. Arrows show differences between the rankings for the two reference periods. The bars in the scatter plots indicate the 10% and the 90% quantiles for IQD values at individual grid points. Colour indicates climate model.

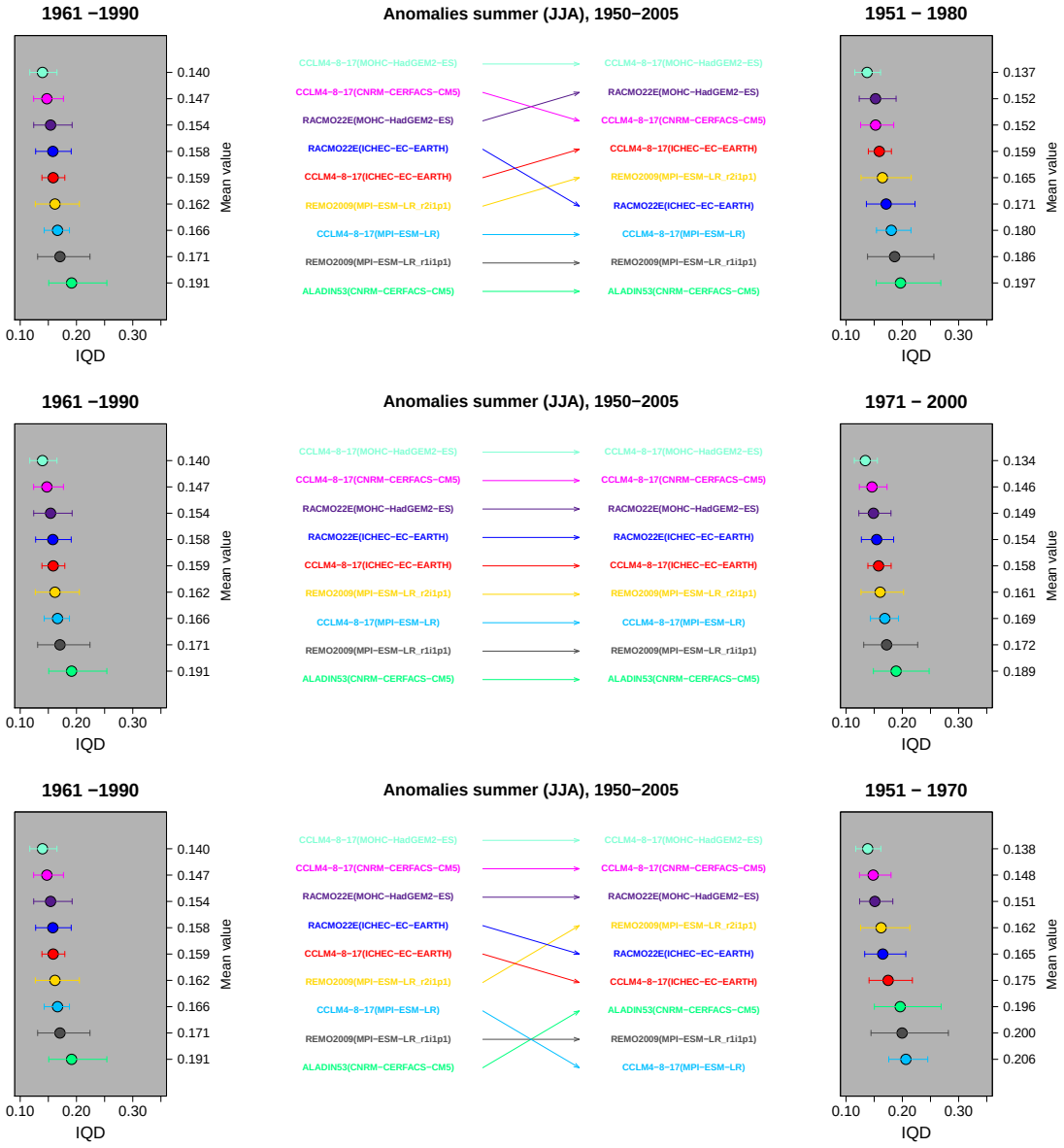


Figure A.1. Summer evaluation: Alternative reference periods on the right are 1951-1980, 1971-2000 and 1951-1970.

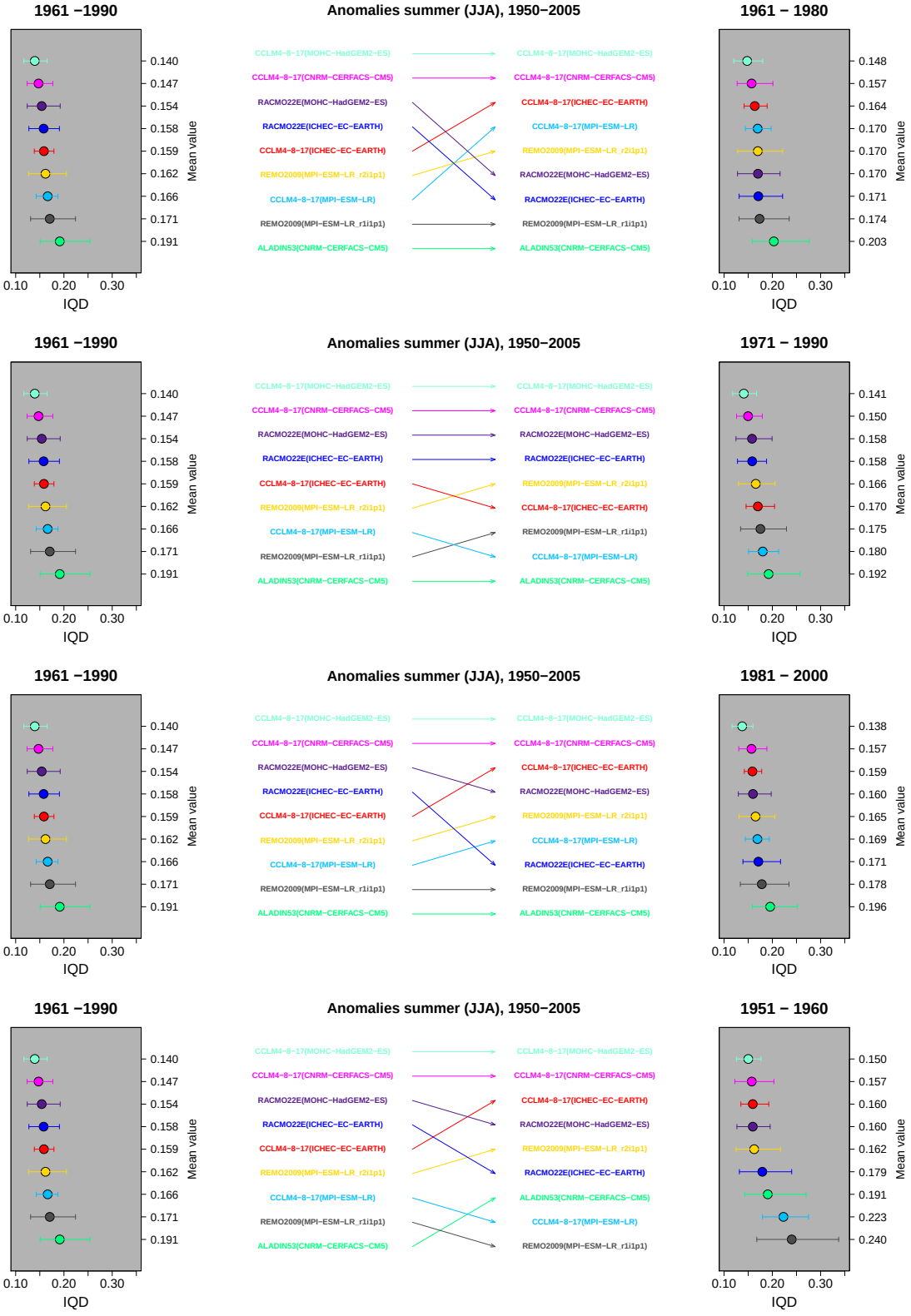


Figure A.2. Summer evaluation: Alternative reference periods on the right are 1961-1980, 1971-1990, 1981-2000 and 1951-1960.

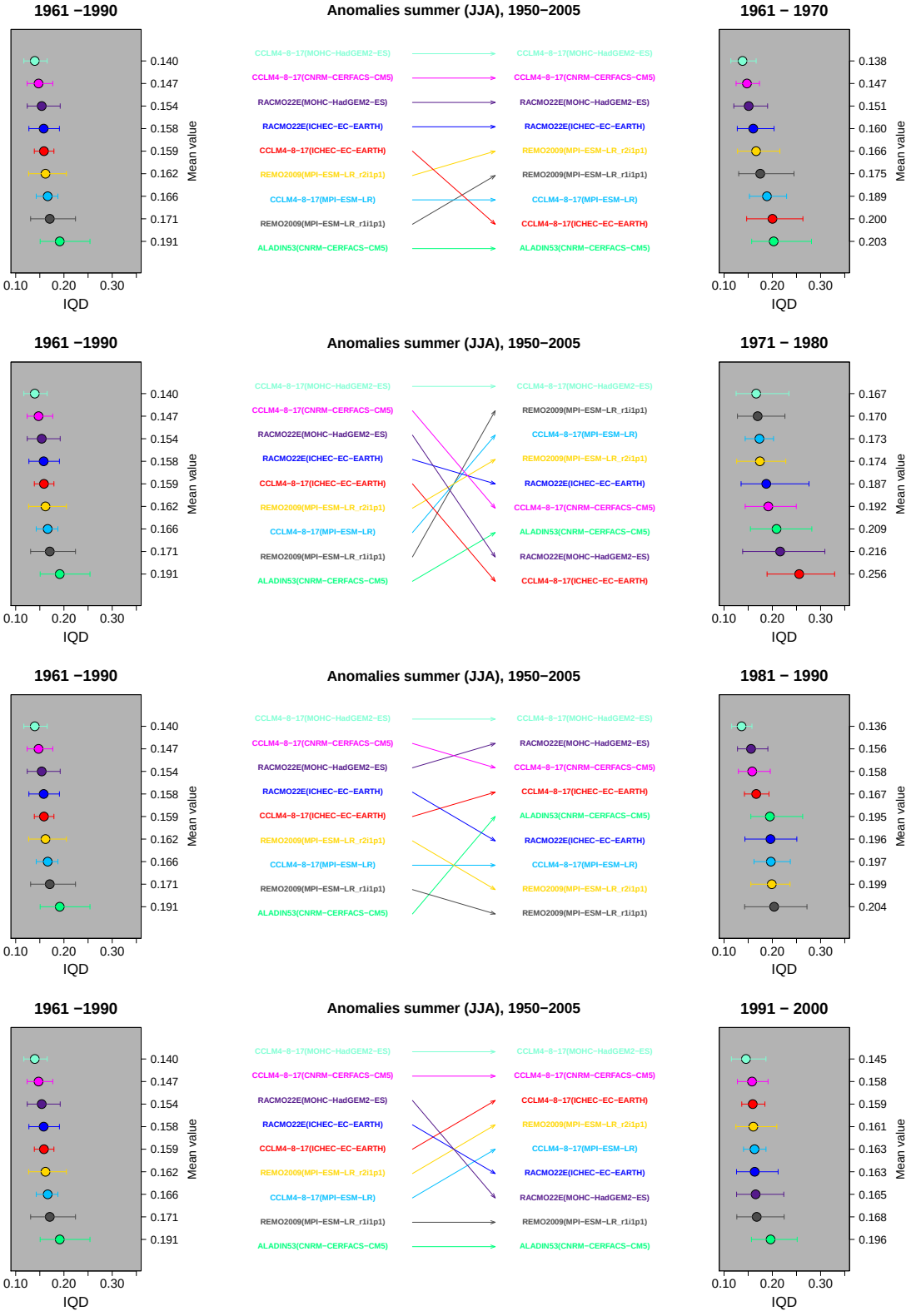


Figure A.3. Summer evaluation: Alternative reference periods on the right are 1961-1970, 1971-1980, 1981-1990 and 1991-2000.

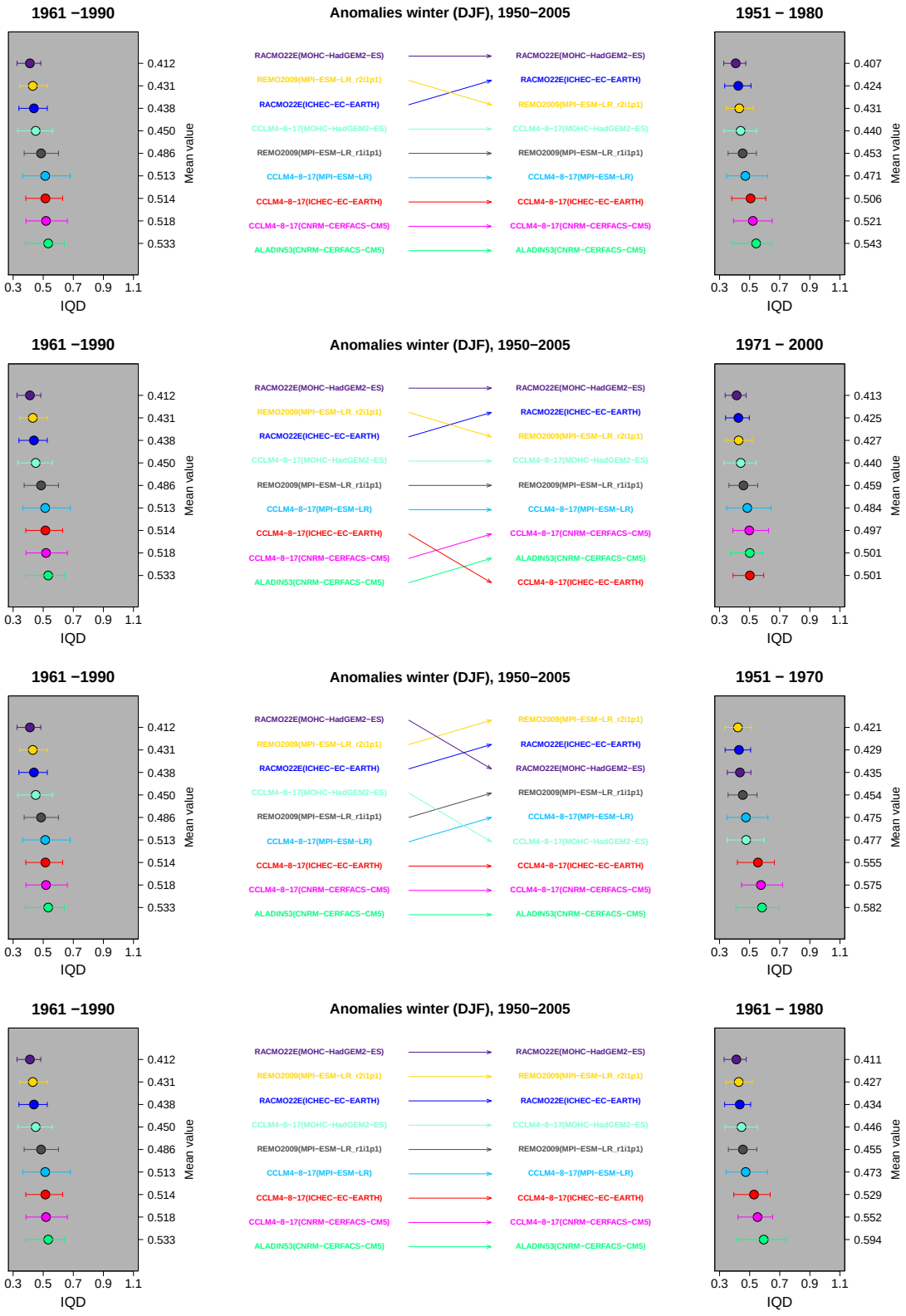


Figure A.4. Winter evaluation: Alternative reference periods on the right are 1951-1980, 1971-2000, 1951-1970 and 1961-1980.

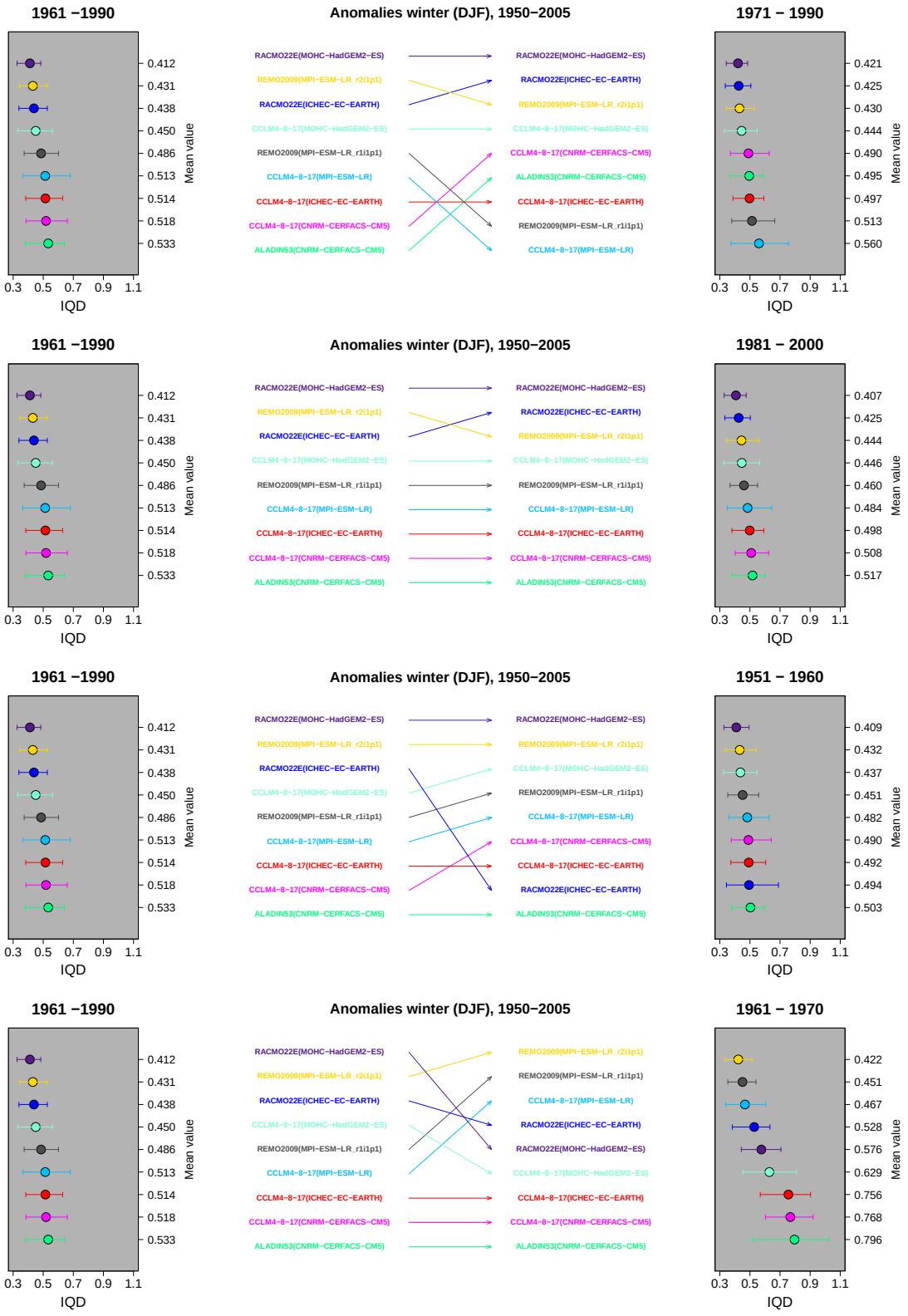


Figure A.5. Winter evaluation: Alternative reference periods on the right are 1971-1990, 1981-2000, 1951-1960 and 1961-1970.

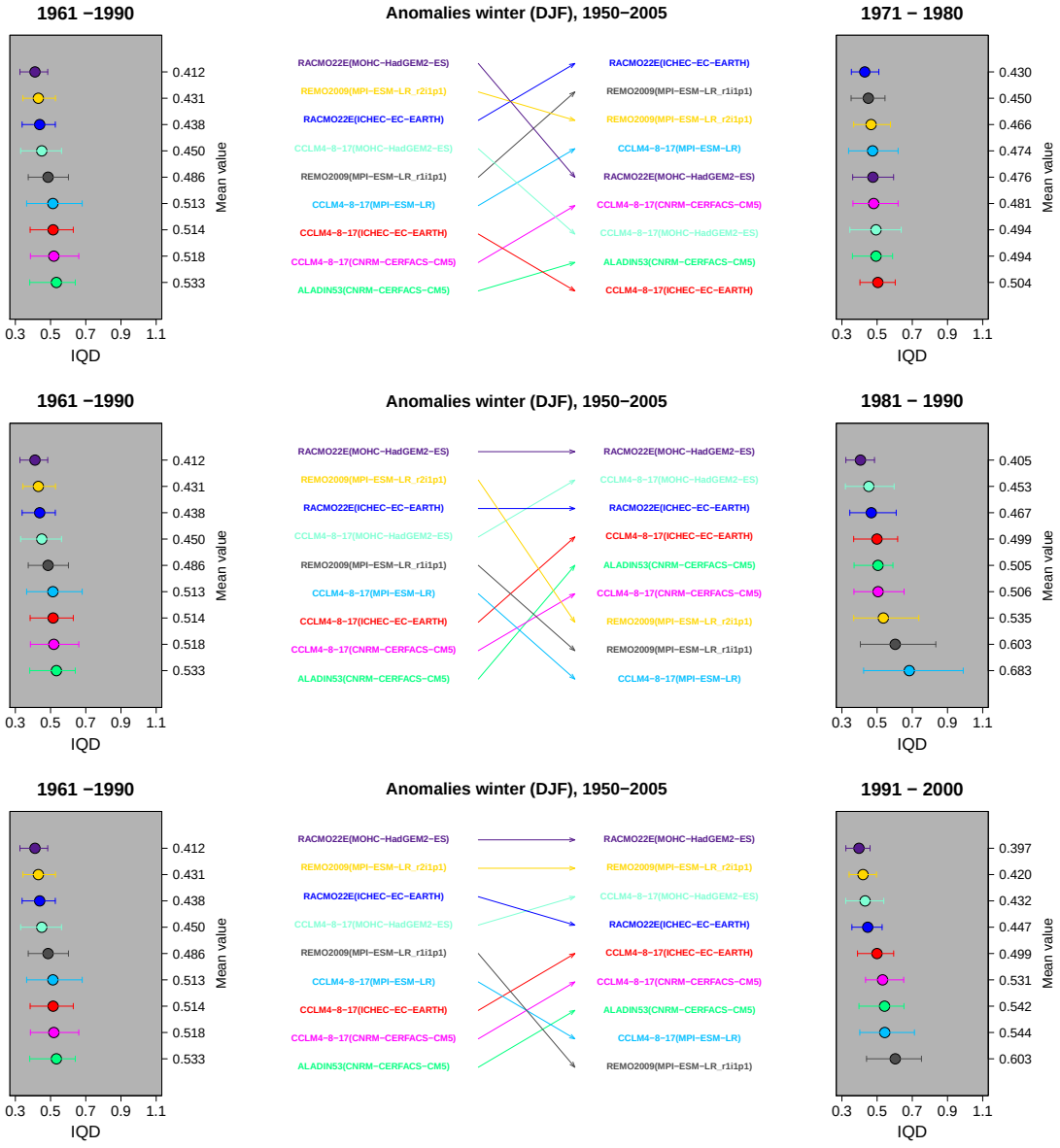


Figure A.6. Winter evaluation: Alternative reference periods on the right are 1971-1980, 1981-1990 and 1991-2000.

B Mean IQD in the summer season

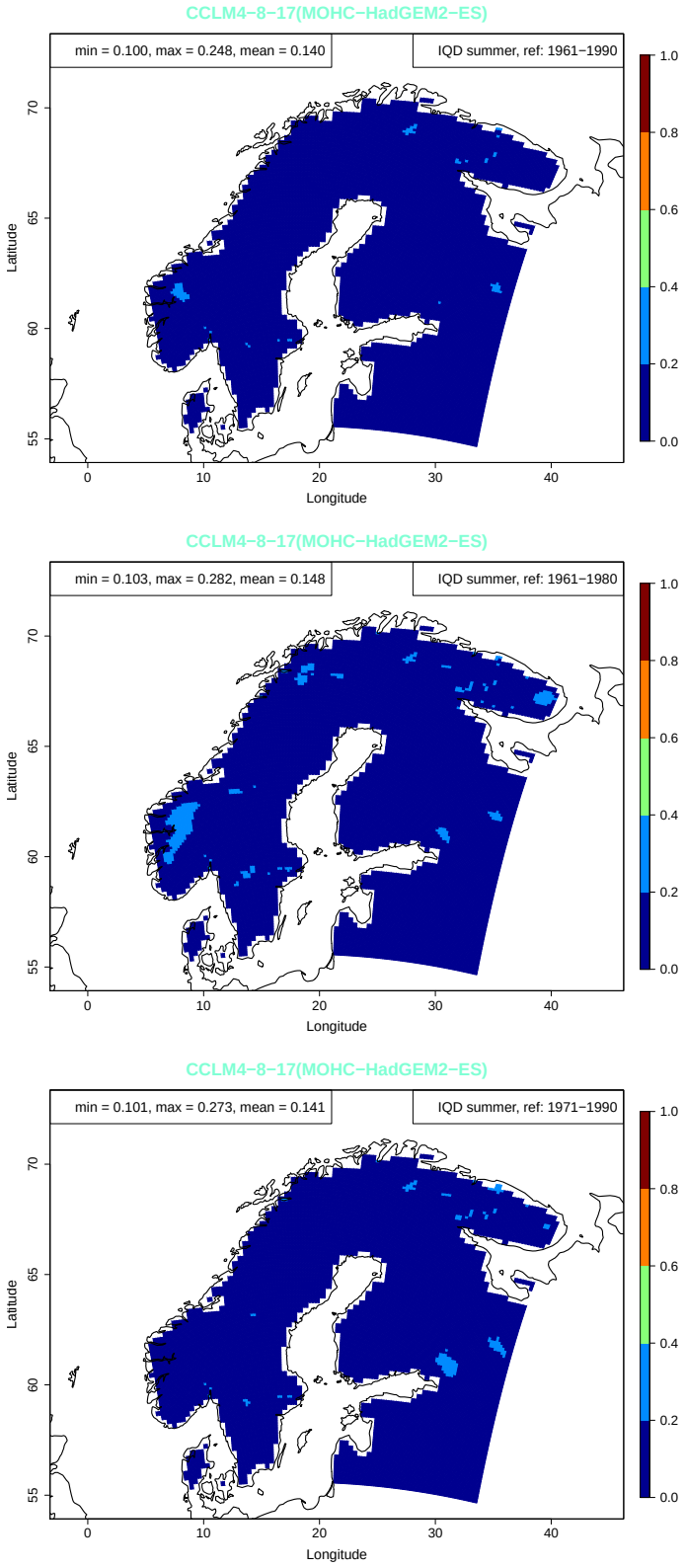


Figure B.7. Mean IQD for the climate model with best performance in the summer season, after applying three different reference periods (1961-1990, 1961-1980 and 1971-1990).

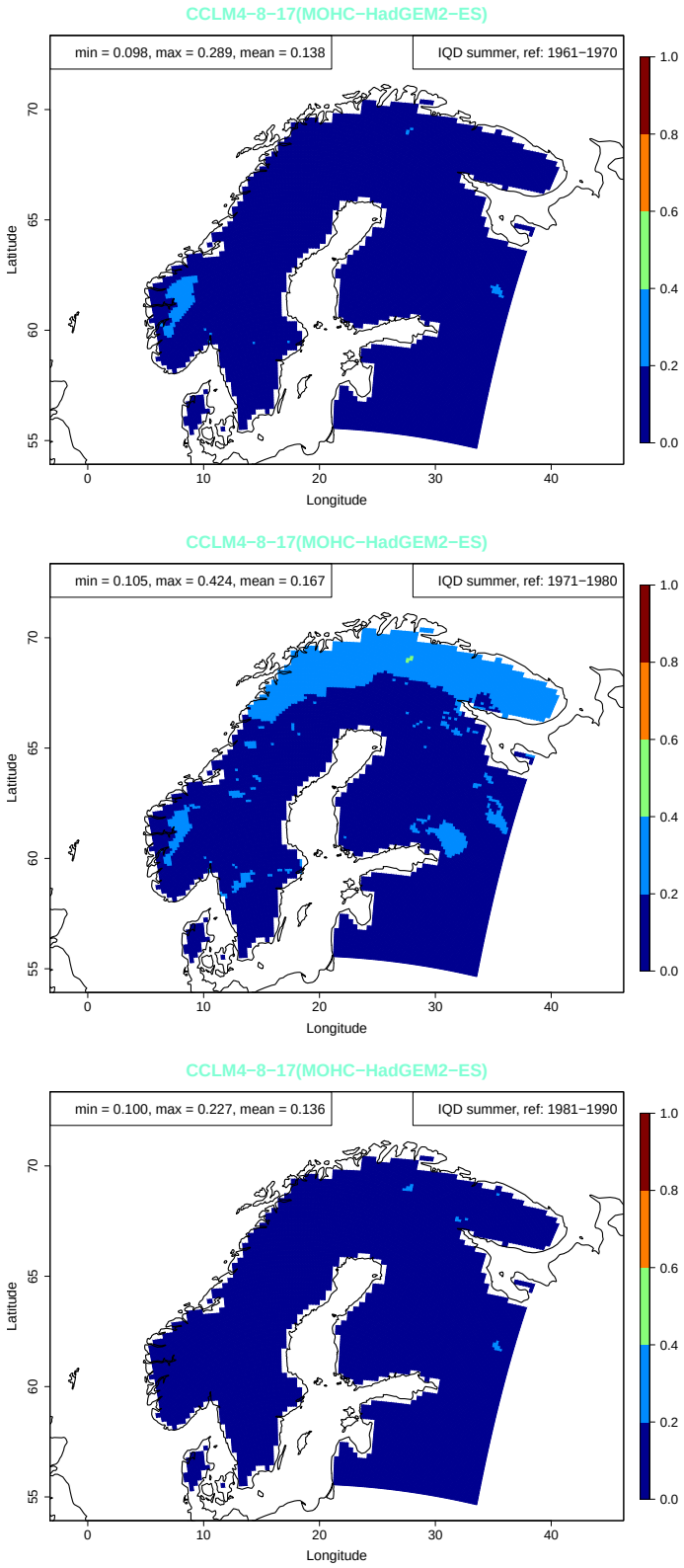


Figure B.8. Mean IQD for the climate model with best performance in the summer season, after applying three different reference periods (1961-1970, 1971-1980 and 1981-1990).

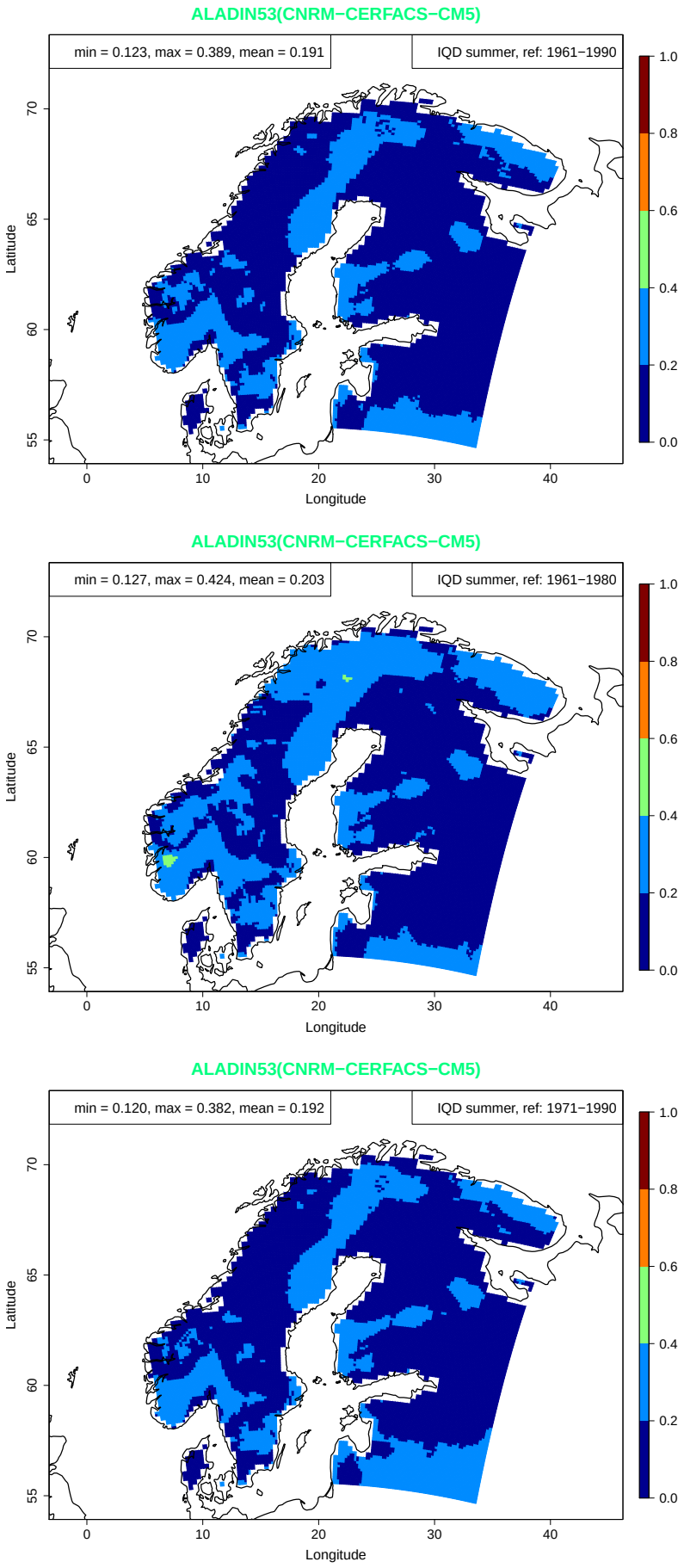


Figure B.9. Mean IQD for the climate model with worst performance in the summer season, after applying three different reference periods (1961-1990, 1961-1980 and 1971-1990).

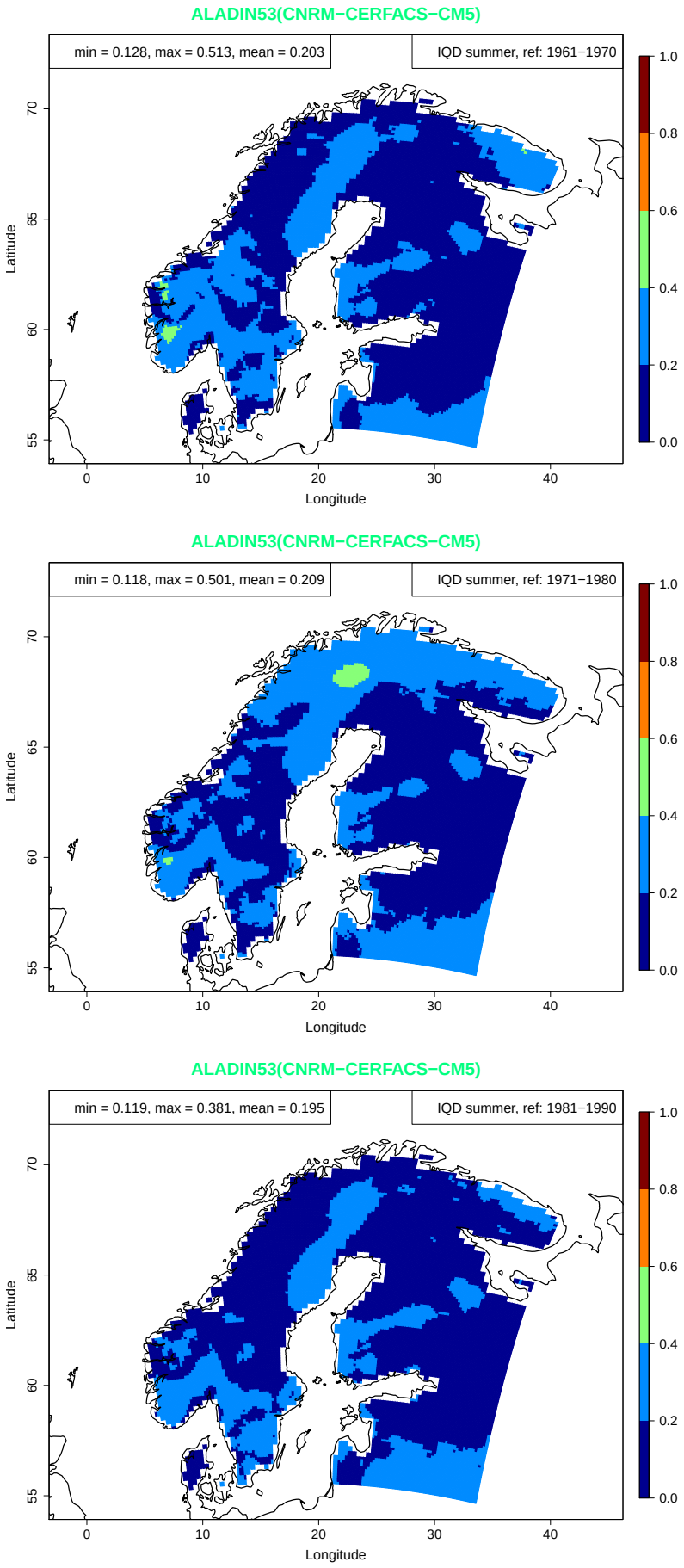


Figure B.10. Mean IQD for the climate model with worst performance in the summer season, after applying three different reference periods (1961-1970, 1971-1980 and 1981-1990).

C Mean IQD in the winter season

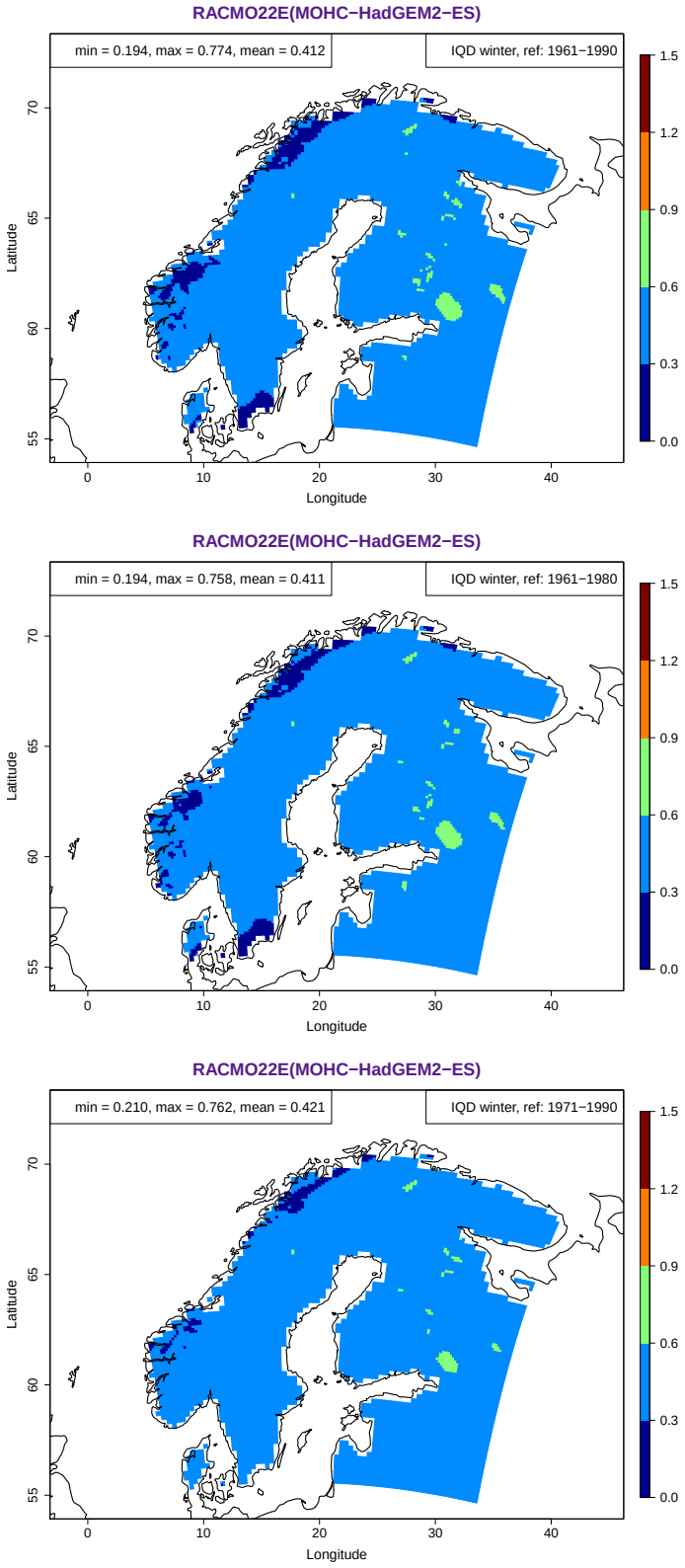


Figure C.11. Mean IQD for the climate model with best performance in the winter season, after applying three different reference periods (1961-1990, 1961-1980 and 1971-1990).

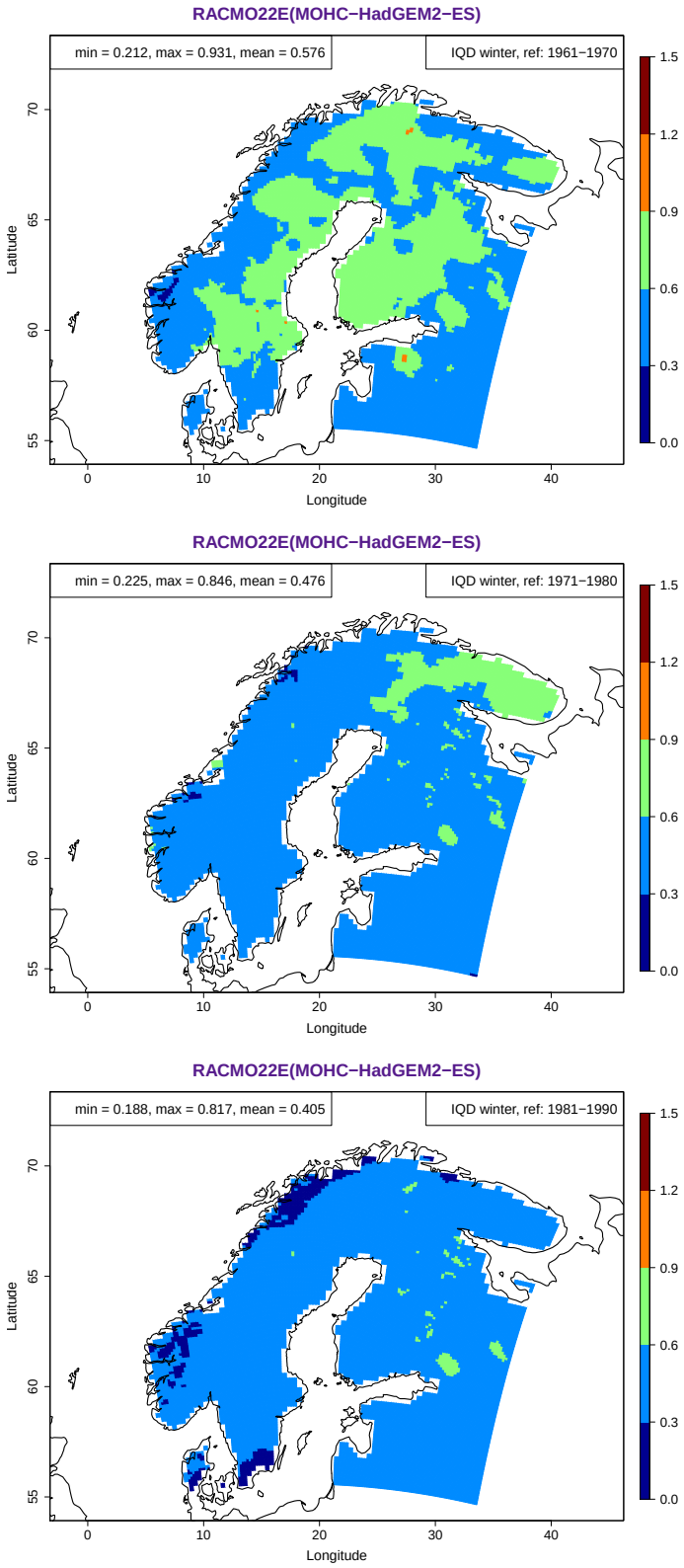


Figure C.12. Mean IQD for the climate model with best performance in the winter season, after applying three different reference periods (1961-1970, 1971-1980 and 1981-1990).

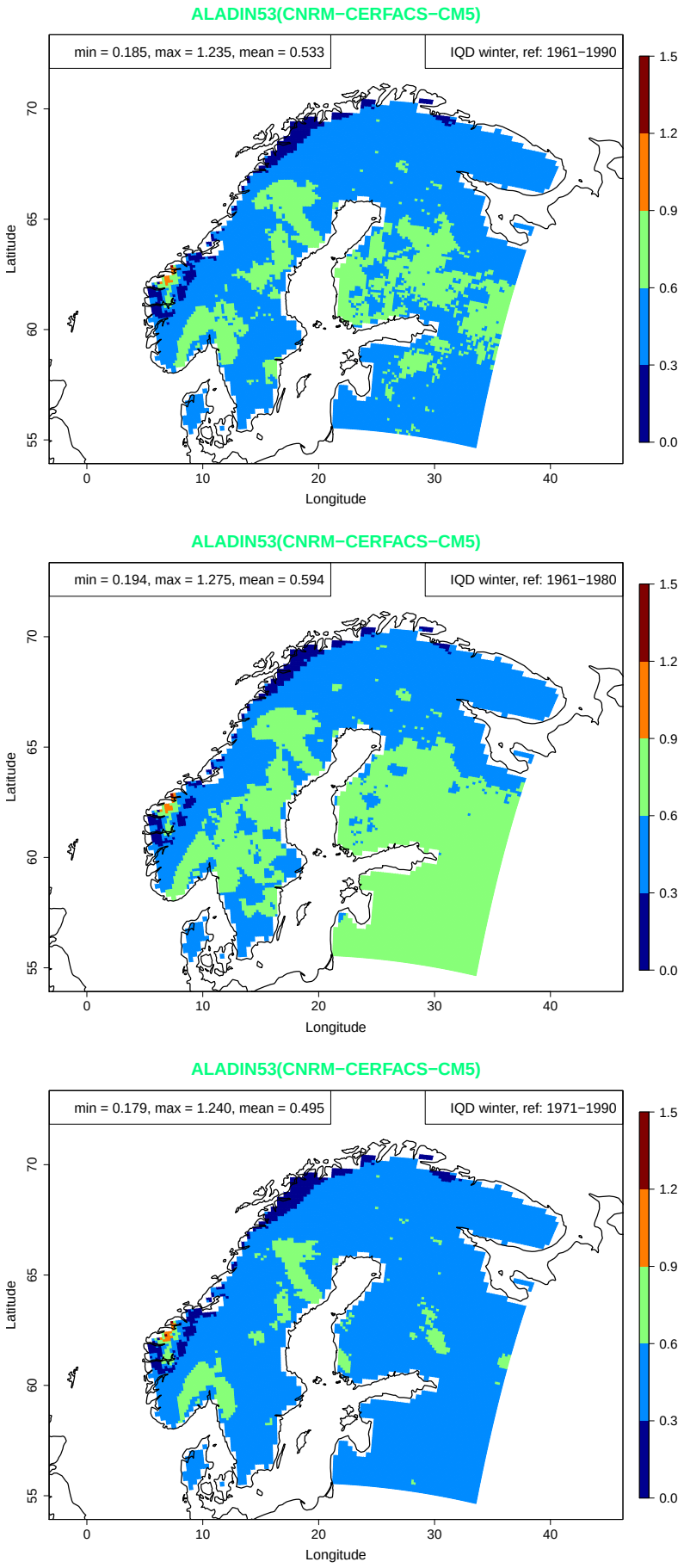


Figure C.13. Mean IQD for the climate model with worst performance in the winter season, after applying three different reference periods (1961-1990, 1961-1980 and 1971-1990).

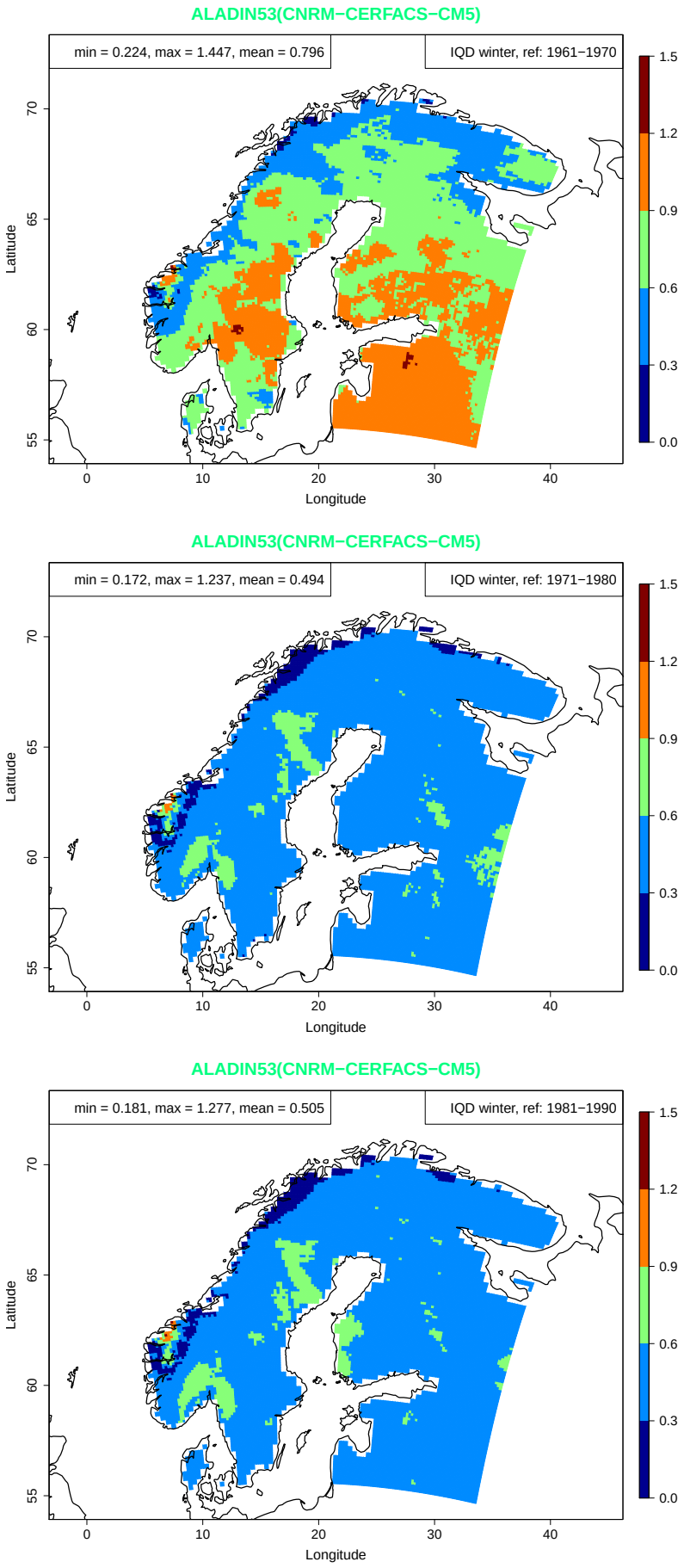


Figure C.14. Mean IQD for the climate model with worst performance in the winter season, after applying three different reference periods (1961-1970, 1971-1980 and 1981-1990).