

Université du Québec
Institut national de la recherche scientifique
Centre Eau Terre Environnement

**DEFINING A REFERENCE CLIMATE FOR CANADA THROUGH THE
COMBINATION OF OBSERVATIONAL AND REANALYSIS DATASETS**

Par
Alexis Pérez Bello

Mémoire présenté pour l'obtention du grade de
Maître es Sciences, M.Sc.
en sciences de l'eau

Jury d'évaluation

Examineur externe	Marie-Amélie Boucher Université de Sherbrooke
Examineur interne	Sophie Duchesne INRS, Centre Eau Terre Environnement
Directeur de recherche	Alain Mailhot INRS, Centre Eau Terre Environnement

Acknowledgments

I want to thank all those people who in one way or another gave me their unconditional support in the realization of this work;

- To my family who have always given me all the necessary support and more.
- To my supervisor Alain Mailhot for the trust, his support, his right guidance, his patience (especially that) and for transmitting his experience in the process of developing this master project.
- To Guillaume Talbot for the constant and huge help from the first day I started. To Dikra Khedhaouiria for your help and advice.
- To all the members of the jury for agreeing to judge my work.

Sincerely, thank you very much

Résumé

Le présent travail propose une combinaison d'ensemble de différents jeux de données à l'aide de deux méthodes d'assimilation pour améliorer la qualité des champs de précipitation climatiques au Canada pour une période de 30 ans (1980-2009). Quatre ensembles de données de réanalyse ont été utilisés : CFSR, ERAI, JRA55 et MERRA. Les données d'observation proviennent de 2160 stations météorologiques à travers le pays avec un minimum de 10 années valides de précipitations quotidiennes. De plus, les données d'observation sur grille de Ressources Naturelles Canada (NRCan pour Natural Resources Canada) ont été utilisées à fin de comparaison. Les Indices de Précipitation Climatiques (IPC) ont été calculés pour chaque ensemble de données annuellement et ont été combinés (la Réanalyse + les Observations) par les méthodes d'Interpolation Optimale (OI pour Optimal Interpolation) et d'Interpolation Optimale d'Ensemble (EnOI pour Ensemble Optimal Interpolation) respectivement. Une vérification a été effectuée sur les stations qui ont été utilisées pour la combinaison, de même qu'une stratégie de validation croisée. Les résultats ont montré des améliorations au niveau des nouveaux ensembles de données de référence créés par les deux méthodes, en particulier pour l'approche d'ensemble. L'ensemble de données sur grille développé (par EnOI) a surpassé NRCan, lequel a été généré seulement à partir d'observations, ce qui montre que ce dernier devrait être utilisé avec précaution pour l'analyse d'événements de précipitation extrême. L'ensemble de données produit pourrait être utilisé pour les études d'impact de changement climatique et de stratégies d'adaptation de même que pour fournir une meilleure compréhension du climat actuel au Canada.

Mots-clés Interpolation optimale d'ensemble, Indices de Précipitations Climatiques, Assimilation de Données

Abstract

The present work shows a combination of different datasets through two data assimilation methods in order to improve the quality of the climate precipitation fields in Canada for a period of 30 years (1980-2009). Four reanalysis datasets were used; CFSR, ERAI, JRA55 and MERRA. The observational dataset consists in 2160 meteorological stations across the country with a minimum of 10 valid years of precipitation records available. Besides, the Natural Resources Canada gridded dataset (NRCan) was also used. The Climate Precipitation Indices were calculated for each dataset annually and were combined (Reanalysis + Observations) by the Optimal Interpolation (OI) and Ensemble Optimal Interpolation (EnOI) methods respectively. The verification was carried out over the stations that were used in the combination, but also through a cross-validation strategy. Results show the improvements in the new reference datasets created by both methods highlighting the Ensemble approach as the best. The reference gridded dataset developed (by EnOI) outperformed NRCan which was created only by observations and showing that the latter should be used with caution for extreme precipitation events analysis. The dataset produced contributes to the development of climate change impact studies and adaptation strategies, as well as providing a better understanding of the current climate in Canada.

Keywords Ensemble Optimal Interpolation, Climate Precipitations Indices, Data Assimilation

Contents

Acknowledgments	iii
Résumé	v
Abstract	vii
Contents	ix
List of Figures	xi
List of Tables	xiii
List of abbreviations	xv
I Synthesis	1
1 Introduction	3
2 Basic notions	5
2.1 Extreme precipitation events and indices	5
2.2 Reanalysis	6
2.3 Optimal Interpolation and Ensemble Optimal Interpolation	7
3 Methodology	9
3.1 Datasets used	9
3.2 Climate Precipitation Indices	11
3.3 Workflow diagrams	11
3.4 Validation procedure	12
4 Results and discussion	15
5 Summary and Conclusions	21
II Article	23
1 Defining a reference climate for Canada through the combination of observational and reanalysis datasets	25
1.1 Abstract	26

1.2	Introduction	27
1.3	Datasets	29
1.4	Climate Precipitation Indices and Super-Observations	31
1.5	Optimal Interpolation (OI)	34
1.6	Ensemble Optimal Interpolation (EnOI)	36
1.7	Cross-validation procedure	37
1.8	Results and Discussion	40
1.8.1	Comparison of reanalysis, OI, and EnOI datasets	40
1.8.2	Cross-validation	42
1.8.3	Comparison of EnOI and NRCan gridded interpolated datasets	46
1.9	Summary and conclusions	50
	References	53

List of Figures

3.1	Creation of super-observations. Panel on left shows the original stations network while the panel on right-hand side shows the corresponding network of super-observations.	10
3.2	Example of the grids for the various reanalysis, blue lines: CFSR; green lines: ERAI; dashed red lines: JRA55; magenta lines: MERRA.	11
3.3	Workflow diagram for OI method	12
3.4	Workflow diagram for EnOI method	12
4.1	Violin and box plots showing the NE (averaged over all ten indices) after the application of the data assimilation techniques. Boxes represent inter-quartiles range (first, Q1 and third, Q3), line inside the box the median value and outliers correspond to the black dots. Outliers are defined as values 1.5 times the interquartile distance ($IQR=Q3-Q1$) (whiskers) above or below Q3 or Q1 respectively.	16
4.2	Annual values of V (see equation 4.1) after applying OI to each reanalysis (top 12 rows) and after applying EnOI (bottom four rows) averaged over all CPI for the various cross-validation subnetworks (<i>Near</i> and <i>Far</i> configurations with 10, 30 and 50% validation sites). Black dots correspond to the datasets with the largest improvements for a given cross-validation subnetwork.	17
4.3	Violin plots showing the bias between EnOI or NRCan and observed values at all the stations for PRmean, PR1mm, RX1day, RX5day, R10mm and R20mm. Boxes represent inter-quartiles range (first, Q1 and third, Q3), line inside the box the median value and outliers correspond to the black dots. Outliers are defined as values 1.5 times the interquartile distance ($IQR=Q3-Q1$) (whiskers) above or below Q3 or Q1 respectively.	18
4.4	Violin plots showing the bias between EnOI or NRCan and observed values at all the stations for CDD, CWD, SDII and PRCPTOT. Boxes represent inter-quartiles range (first, Q1 and third, Q3), line inside the box the median value and outliers correspond to the black dots. Outliers are defined as values 1.5 times the interquartile distance ($IQR=Q3-Q1$) (whiskers) above or below Q3 or Q1 respectively.	19
1.1	Map of the stations considered in this study.	31
1.2	Number and percent (above each bar) of stations with given numbers of valid years.	31
1.3	Total number of stations and super-observations available each year.	33
1.4	Estimated correlation function (continuous curve) with corresponding discrete values for the RX1day from CFSR.	35
1.5	Spatial distribution of calibration (red) and validation (blue) sites on year 1980 for (a) <i>Near</i> and (b) <i>Far</i> subnetworks with 10% validations stations.	39

1.6	Violin and box plots showing the distances between all validation sites and their closest calibration sites for <i>Near</i> and <i>Far</i> subnetworks respectively over the study period. Boxes represent inter-quartiles range (first quartile Q1 and third quartile Q3), line inside the box the median value and outliers correspond to the black dots. Outliers are defined as values 1.5 times the interquartile distance ($IQR=Q3-Q1$) (whiskers) above or below Q3 or Q1 respectively.	39
1.7	Taylor diagrams comparing the observed and estimated annual CPI values over the 1980-2010 period at all station sites: a) before applying OI; b) after applying OI; c) after applying EnOI.	41
1.8	Annual NRMSD (averaged over all stations and all CPI) values normalized by the observational standard deviation (σ over all valid stations for a given year) for each reanalysis after applying OI and EnOI.	42
1.9	(a) Distance between each grid point and the closest station and (b) expected relative CPI error (equation 1.7) as a function of the distance to closest station.	43
1.10	Mean Absolute Error of the analysis field (MAE_A) normalized by the Mean Absolute Error of the background field (MAE_B) over all validation sites for the entire period (1980-2009) (equation 1.14) and each CPI (x-axis). <i>Near</i> (N) and <i>Far</i> (F) configurations with 10, 30, and 50% validation points after applying OI to the four reanalyses (top four rows) and after applying EnOI (bottom row) are presented. Black dots identify, for each CPI, the datasets with the lowest V_1 values for a given subnetwork (N or F with 10, 30, or 50% of validation points).	45
1.11	RMSD of the analysis field ($RMSD_A$) normalized by the standard deviation for each CPI (σ_B) of annual values at all validation sites for the entire period (1980-2009) (equation 1.15) and each CPI (x-axis). <i>Near</i> (N) and <i>Far</i> (F) configurations with 10, 30 and 50% of validation points after applying OI to the four reanalyses (top four rows) and after applying EnOI (bottom row) are presented. Black dots identify the datasets with the best performance (smallest values) for a given subnetwork (N or F with 10, 30, or 50% validation points) and CPI.	46
1.12	Mean PRmean (first line), PR1mm (second line), and RX1day (third line) values over the 1980-2009 period as estimated from the EnOI (left column), NRCAN datasets (middle column), and corresponding grid-to-grid differences between EnOI and NRCAN (right column).	47
1.13	Scatter plots of PRmean (first line), PR1mm (second line), and RX1day (third line) values estimated from EnOI (left column) and NRCAN (right column) dataset against the corresponding observed values.	48
1.14	Taylor diagram comparing annual CPI values at station sites estimated with NRCAN gridded dataset (red symbols) and after applying EnOI (blue symbols).	49

List of Tables

- 1.1 Gridded datasets used in this study (acronyms appearing in parenthesis are used to refer to these datasets). 30
- 1.2 Climate Precipitation Indices, adapted from Zhang *et al.* (2011) 32

List of abbreviations

AHCCD	Adjusted and Homogenized Canadian Climate Data
ANUSPLIN	Program for interpolating data using a thin plate spline smoothing algorithms
CFSR	Climate Forecast System Reanalysis
CMIP5	Coupled Model Intercomparison Project Phase 5
CPI	Climate Precipitation Indices
EnKF	Ensemble Kalman Filter
EnOI	Ensemble Optimal Interpolation
ERA-I	ERA-Interim Reanalysis
ETCCDI	Expert Team on Climate Change Detection and Indices
IPCC	Intergovernmental Panel on Climate Change
JRA55	Japanese 55-year Reanalysis
MAE	Mean Absolute Error
MDDELCC	Ministère du Développement Durable, de l'Environnement et de la Lutte contre les Changements Climatiques
MERRA	Modern-Era Retrospective Analysis for Research and Applications
NE	Normalized Error
NRCan	Natural Resources Canada (observational gridded dataset)
NRMSD	Normalized Root Mean Square Difference
NSD	Normalized Standard Deviation
OI	Optimal Interpolation
RMSD	Root Mean Square Difference

Part I

Synthesis

Chapter 1

Introduction

During the last decades, many major precipitation events have been reported in many parts of the world (Hartmann *et al.*, 2013). These events are expected to be more frequent and intense over the course of the 21st century due to climate change (Stocker *et al.*, 2013). It is therefore important to assess how these types of events will evolve in a future climate in order to better assess their potential impacts on ecosystems, economy and the societies as a whole and to develop adequate adaptation strategies.

Defining and characterizing the historical climate is crucial to assess the projected future changes (Hatzaki *et al.*, 2010). Reference climate can be defined and characterized in many ways. The Expert Team on Climate Change Detection and Indices (ETCCDI) proposed a set of annual precipitation indices that can be estimated from daily precipitation series (Peterson *et al.*, 2001; Zhang *et al.*, 2011) (Climate Precipitation Indices, hereafter CPI). Spatial distribution of such indices over a given region requires a dense network of meteorological observations and historical climate can be readily defined in regions where a large meteorological network has been deployed over the recent decades (Heim, 2015). However, this remains a huge challenge for regions with sparse network or with short records. In such cases, other datasets must be considered. Despite the existence of these alternative datasets, it must be reminded that records from meteorological stations remain the primary and possibly the most reliable source of data for historical climate. The lack of a dense network of observations with long time records, especially in the northern parts of Canada, therefore represents a huge challenge (Wang *et al.*, 2017).

Reanalyses are an interesting alternative source of information for different climate applications. They provide consistent gridded datasets in space and time that could be useful for regions with sparse network or short records. Despite improvements in the spatio-temporal representation of many variables brought by reanalysis, precipitation remains a problematic field because observations of this variable have not been explicitly used to create most of the available reanalyses (Zwiers *et al.*, 2013). However, Diaconescu *et al.* (2017) showed that reanalyses should be used with caution for the northern parts of Canada. Therefore, reanalyses need to be bias corrected before they can be used as a reference dataset. One possible option for improving the CPI representation is through the combination of reanalysis datasets with *in situ* observations through data assimilation methods.

The aim of this study is to improve the quality of the annual climate precipitation fields for Canada over the period 1980-2009. It is based on the hypothesis that it can be done by combining annual CPI from surface weather stations with reanalyses through data assimilation techniques. In order to merge both datasets, Optimal Interpolation (hereafter, OI) and Ensemble Optimal Interpolation (hereafter, EnOI) methods were used. This document is divided in two parts. The first part (“Synthesis”) is a summarized description of the study as well as complementary information (Basic notions, Methodology, Results, Conclusions) that provides details about the different steps used in this study. The second part is a scientific paper ready for submission.

Chapter 2

Basic notions

2.1 Extreme precipitation events and indices

The Intergovernmental Panel on Climate Change (IPCC), Special Report on Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation (Seneviratne *et al.*, 2012) defines an *extreme climate event* or an *extreme weather event* as “*the occurrence of a value of a weather or climate variable above (or below) a threshold value near the upper (or lower) ends of the range of observed values of the variable*”. For precipitation values, this definition basically refers to events located in the tails of the probability distributions with low frequency of occurrence. Extreme events can also be analyzed in many different ways depending on applications (Zwiers *et al.*, 2013). For instance, for low frequency of occurrence (more than a year), different methods based on extreme value theory (Coles, 2001) can be used. Those methods are based on statistical models designed to estimate the probability of occurrence of such events over a period beyond the recorded one (Zhang & Zwiers, 2013). Some examples of application of those methods for extreme precipitation events are presented in (Kharin *et al.*, 2007; DeGaetano, 2009; Mishra & Singh, 2010; Rajczak *et al.*, 2013; Rahimpour *et al.*, 2016; Santos *et al.*, 2016).

The Expert Team on Climate Change Detection and Indices (ETCCDI) proposed a set of indices to characterize extreme precipitation and temperature (only precipitation indices are considered hereafter). A majority of these indices aim at characterizing events that occur at least once a year and are called “*moderate extremes*” (Zhang *et al.*, 2011). These indices can be divided in three

groups depending of their characteristics. The first group includes indices that are absolute values of the variable analyzed. The Annual maximum of daily precipitation (RX1day) as well as the annual maximum precipitation over 5 consecutive days (RX5day) belong to this group. These indices can be used to estimate annual maximum values of different return periods. The second group includes indices that meet a given condition, defined by a fixed threshold, counting the number of days when this condition is met. Example of such indices are the number of days with precipitation values with more than 10, 20 or 30 mm (R10mm, R20mm, R30mm) and the number of wet days (days with precipitation values above 1 mm, PR1mm). This group also includes indices related to persistence of given meteorological conditions such as the annual maximum number of consecutive dry days (when precipitation is below 1 mm, CDD) and the annual maximum number of consecutive wet days (when precipitation is above 1 mm, CWD). The third group includes indices that refer to extreme quantiles of the index distribution over a specified climatological period. An example of such index is the annual number of days with daily precipitation above the 95th percentile of precipitation distribution defined over a specific reference period. This third group of indices were not considered in this study.

Extreme precipitation indices have been used to characterize changes in frequency and intensity in different regions around the world. Five precipitation indices over the period 1946-1999 were analyzed by Frich *et al.* (2002), and global maps representing multi-decadal changes at observational sites were created. In that study, despite Africa and South America were not considered, resulting analysis showed significant increases in climate extremes. A comprehensive and updated global gridded land-based representation of extreme precipitation indices was presented by Alexander *et al.* (2006) using eleven indices for the period 1951-2003, and later by Donat *et al.* (2013), using 12 indices over the period 1901-2010. In both studies, detected trends indicated an increase toward wetter conditions in many regions of the globe.

2.2 Reanalysis

Over the last decades, reanalyses have been increasingly useful for many climate applications. A considerable amount of research has been realized using these datasets. Reanalyses are created from the combination of a numerical model and different sources of observational datasets through a data assimilation routine (Bosilovich *et al.*, 2008). They provide a consistent and comprehensive

three dimensional grid representation of the main atmospheric variables throughout the world (Dee *et al.*, 2011). Their quality have been improved recently due to the progress reached in data assimilation methods, the rescue of past observations as well as the homogeneous products obtained from satellites (Kobayashi *et al.*, 2015). Their performance depends of different factors, like: the data assimilation technique, the variable considered for the analysis, the sources of observations used, as well as the forecast model (Lorenz & Kunstmann, 2012). As these factors have been continuously improving over the last years, it is evident that many reanalyses should be considered for specific applications (Saha *et al.*, 2010). The extreme precipitation indices derived from reanalyses have been used in different studies. Sillmann *et al.* (2013) used four reanalysis data sets to assess the performance of Coupled Model Intercomparison Project Phase 5 (CMIP5) models using 11 precipitation indices. Reanalyses were also compared to global gridded observational datasets by Donat *et al.* (2014) using 2 precipitation indices. These authors showed the spatial distribution of precipitation as simulated by reanalyses improved in the satellite era. For the northern parts of Canada, Diaconescu *et al.* (2017), considering 10 precipitations indices, showed that reanalyses should be used with caution over these regions for precipitations related indices. It is important also to note that precipitation is a direct output of numerical models and that ground precipitation records are not assimilated (Kalnay *et al.*, 1996).

2.3 Optimal Interpolation and Ensemble Optimal Interpolation

The OI method (Kalnay, 2002), also known as *statistical interpolation* (Daley, 1991), combines the information from two datasets (e.g. simulated and observed datasets) by exploiting the horizontal correlation distribution between these datasets. The resulting field is expected to be a better representation of the real conditions. EnOI (Evensen, 2003) is an approximation of the Ensemble Kalman Filter (EnKF) in which an ensemble of model states in long time integration (stationary) is used to compute the analysis field (field obtained after the merging process) and only one model state is used as background field (initial field before the merging process) (Oke *et al.*, 2006). Both methods solve the same analysis equation and differ only in how the (background or model) error covariance is defined (Ren & Hartnett, 2017). While for OI it can be created depending on the distance between the stations and can be represented as a continuous function (e.g. such functions can be found in Buell (1972)), for EnOI an ensemble of model states is needed.

OI method has been used to merge different sources of precipitation datasets. A system for mesoscale analyses in the north of Europe was developed by Häggmark *et al.* (2000) where different sources of observations (e.g. satellites, radars and automatic stations) were used. Over Canada, a preliminary real-time precipitation analysis at 15 km of spatial resolution and 6 hours of precipitation accumulation is described in Mahfouf *et al.* (2007). An evaluation of this dataset was later provided by Lespinas *et al.* (2015) and the assimilation of precipitation estimations obtained from radars were added by Fortin *et al.* (2015). For climatological purposes, a 5.5 km grid spacing precipitation reanalysis over Europe was created by Soci *et al.* (2016). Furthermore, for global scale, satellites information, surface observations and models were combined with significant improvements also at regional scales (Nie *et al.*, 2016).

These methods have been also used to assimilate or combine other variables from the ocean-atmosphere system. Some examples in ocean data assimilation can be found in (Kimoto *et al.*, 1997; Oke *et al.*, 2010; Fu *et al.*, 2011; Ratheesh *et al.*, 2014; Ji *et al.*, 2015; Mignac *et al.*, 2015). They have been also applied to assimilate some wave parameters improving the quality of the forecasts, (Qi & Cao, 2015; Cao *et al.*, 2015; Rusu & Guedes Soares, 2016). Furthermore, for air quality analysis (Candiani *et al.*, 2013; Tang *et al.*, 2015) and for daily streamflow (retrospective estimation) in ungauged rivers by Lachance-Cloutier *et al.* (2017). These widely used methods were used in this work to combine reanalysis datasets with surface weather stations over Canada.

Chapter 3

Methodology

This chapter intends to provide complementary information to help understand the different steps of the methodology used in this study. The main steps are presented in the paper (See Article 1).

3.1 Datasets used

Figure 1.1 of the paper presents a map of the ground weather stations used in this study. Daily precipitation records from the Adjusted and Homogenized Canadian Climate Data (AHCCD) developed by Environment Canada and from the stations operated by *Ministère du Développement Durable, de l'Environnement et de la Lutte contre les Changements Climatiques* (MDDELCC) from Quebec were used. Only stations with available records during the period 1980-2009 were considered. As can be seen, stations unevenly cover the territory and are mainly concentrated in the southern part of the country. Another important limitation is the significant number of missing values in the available records (Figure 1.2). Table 1.1 of the paper provides a summary of the characteristics of the gridded datasets used in this study. Those are divided into two groups. The first group comprises four reanalyses (CFSR, ERAI, JRA55, and MERRA) and the second group includes the NRCan observational gridded dataset. It is important to mention that not all the stations considered in this study were used to create the NRCan dataset (for instance, some stations operated by the MDDELCC in Quebec were not used). This dataset is available for the whole territory of Canada and it is easily accessible. In order to compare reanalysis grid points

(interpreted as average values over a grid cells) and stations (at-site value), super-observations or upscaled observations (Cherubini *et al.*, 2002) were defined for grid cells containing many stations. Figure 3.1 shows a graphical representation of how super-observations were created and Equation 1.1 of the paper shows how their values were estimated. Reanalysis grid points (in this case CFSR) are represented by the black dots and the dashed lines define the grid cells, red dots represent the stations while super-observations are represented by green dots (position of super-observations are determined by the weighted centroid of stations positions; the position of super-observations for grid cells with one station are superimposed to station position). Figure 1.3 of the article shows the number of observations per year over the study period as well as the number of super-observations, the latter were chosen to combine with reanalysis by OI and EnOI methods respectively.

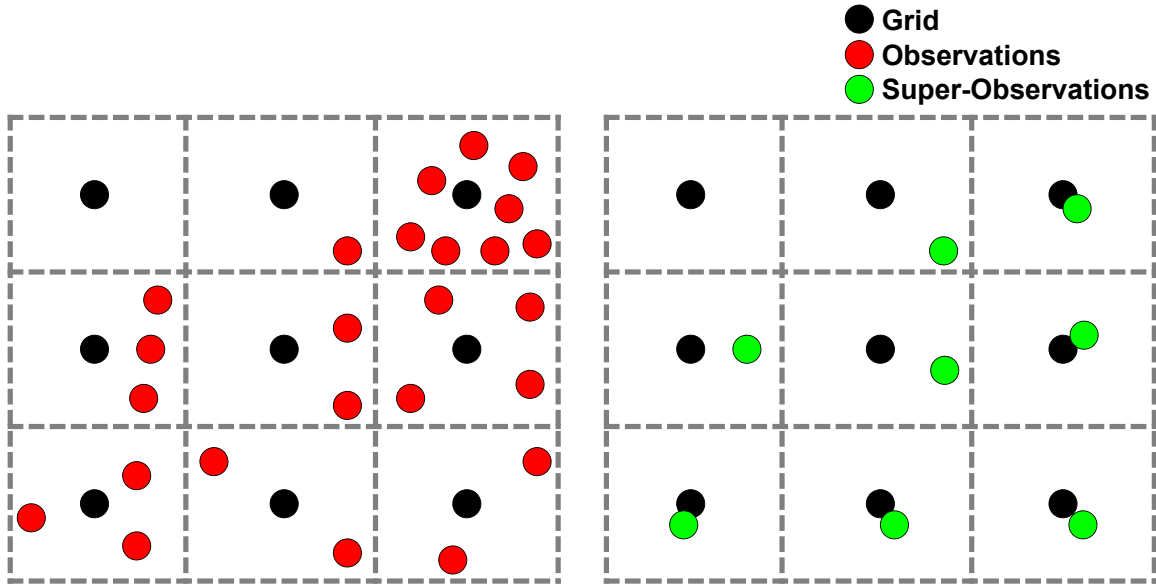


Figure 3.1 – Creation of super-observations. Panel on left shows the original stations network while the panel on right-hand side shows the corresponding network of super-observations.

Table 1.1 of the paper shows that reanalysis grids have different spatial resolution. Figure 3.2 presents an example of each of these grids.

In order to apply the EnOI method which combines the information from the four reanalyses, a common grid must be defined. The grid of the reanalysis with the finest resolution, CFSR, was selected. Once CPI values are estimated from each reanalysis on their respective native grid, a conservative remapping procedure (Jones, 1999) was used to match CPI values from each native grid to the reference grid.

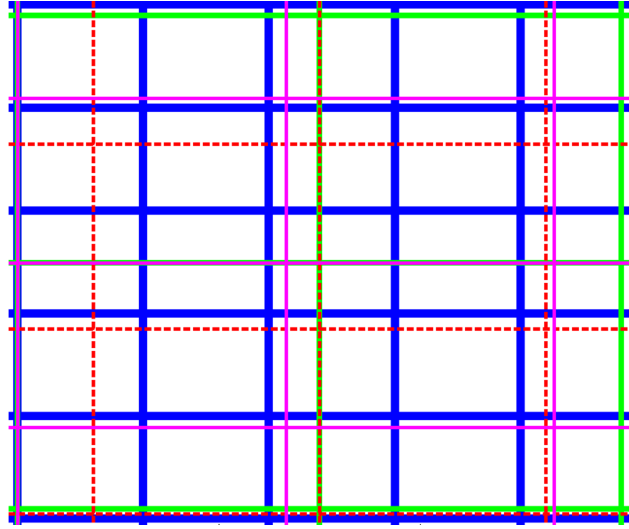


Figure 3.2 – Example of the grids for the various reanalysis, blue lines: CFSR; green lines: ERAI; dashed red lines: JRA55; magenta lines: MERRA.

3.2 Climate Precipitation Indices

The ETCCDI (Peterson *et al.*, 2001; Zhang *et al.*, 2011) has defined a set of temperature and precipitation extreme indices. The list of CPI considered in this study is presented in Table 1.2 of the article. In addition to the selected ETCCDI indices, the annual mean of daily precipitation (PRmean) was also estimated. These indices were calculated for each dataset (reanalyses, NRCan and observations) from daily precipitation values at each grid point of their respective native grid.

3.3 Workflow diagrams

Figure 3.3 and 3.4 shows the workflow diagrams for the OI and EnOI methods respectively. EnOI uses the information provided by the four reanalyses together and the initial field used was the ensemble mean while OI was applied for each reanalysis independently. Details about these two approaches are provided in section 1.5 and 1.6 of the article. These methods were applied independently each year rather than over the climatological mean over the 30-year study period. The main reason justifying this procedure was to make the most of available records.

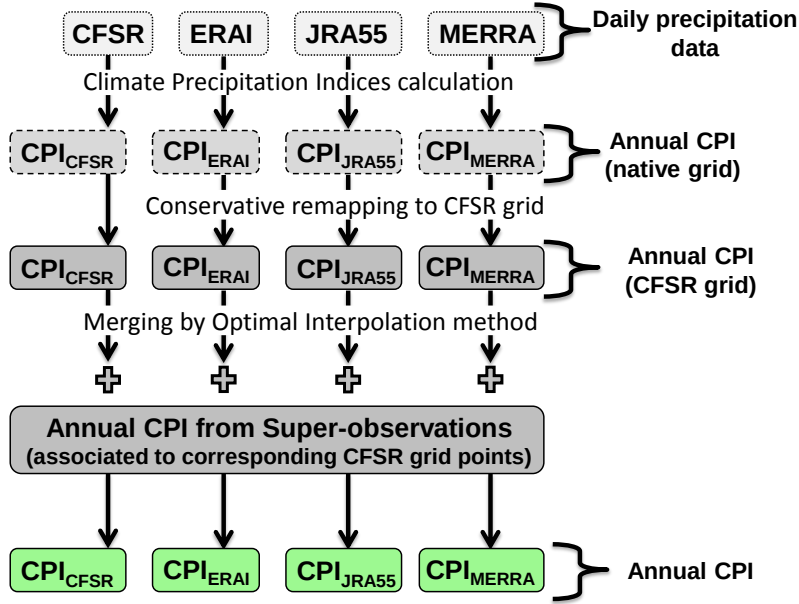


Figure 3.3 – Workflow diagram for OI method

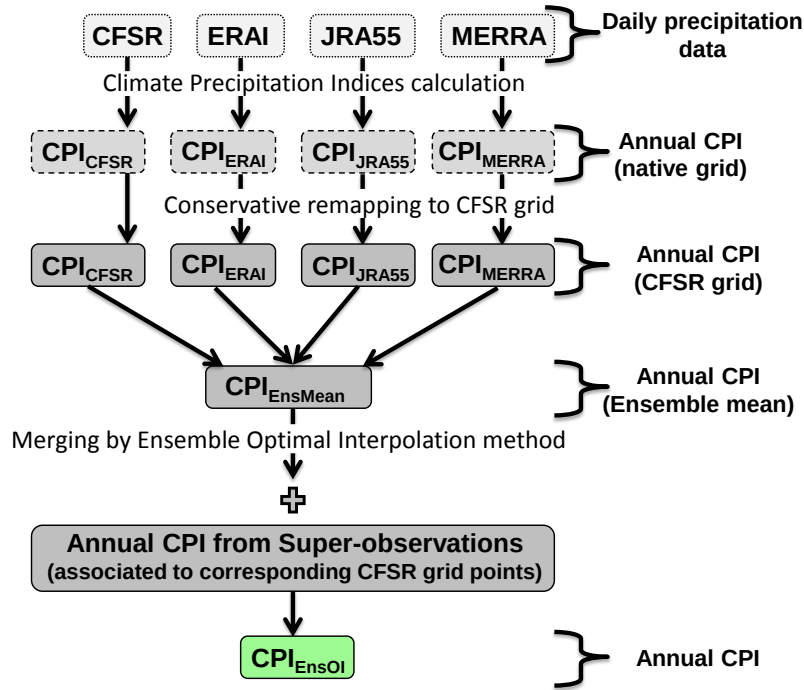


Figure 3.4 – Workflow diagram for EnOI method

3.4 Validation procedure

Preliminary validation of the OI and EnOI datasets was carried out by comparing the estimated CPI values to CPI values estimated from station records. Taylor diagrams (Taylor, 2001) were used

to compare the different CPI before applying OI and EnOI (raw dataset directly from reanalysis) and after their application. A Taylor diagram summarizes the information about the Root Mean Square Difference (RMSD) between two datasets through three elements: the Root Mean Square Difference (RMSD) itself, the correlation coefficient (R), and the standard deviation (σ). In order to compare multiple indices on the same diagram, the normalized version of this diagram was used where the RMSD and the standard deviation of each CPI were normalized by the average standard deviation from the observations. The Normalized RMSD (NRMSD), the Normalized Standard Deviation (NSD) and R are defined by the following equations:

$$NRMSD = \frac{RMSD}{\sigma_o}, \quad NSD = \frac{\sigma_m}{\sigma_o}, \quad R = \frac{\frac{1}{N} \sum_{i=1}^N (M_i - \bar{M})(O_i - \bar{O})}{\sigma_m \sigma_o} \quad (3.1)$$

with

$$RMSD = \sqrt{\frac{1}{N} \sum_{i=1}^N [(M_i - \bar{M}) - (O_i - \bar{O})]^2} \quad (3.2)$$

$$\sigma_m = \sqrt{\frac{1}{N} \sum_{i=1}^N (M_i - \bar{M})^2} \quad \text{and} \quad \sigma_o = \sqrt{\frac{1}{N} \sum_{i=1}^N (O_i - \bar{O})^2} \quad (3.3)$$

where N represents the number of all available super-observations over the study period, M and O represent given model and observation CPI values (at super-observation points) respectively, while $\bar{M}$ and $\bar{O}$ their respective mean values. The standard deviation of model and observations CPI values are defined by σ_m and σ_o respectively.

Another metric was used to compare the performance between the different datasets after the application of OI and EnOI. It was the Normalized Error (NE), which provides information of how good were the estimated CPI relative to the CPI values from the observations. It was applied for each CPI estimated at each super-observation point i and it is defined by the following equation:

$$NE_{(i)} = \frac{M_i - O_i}{O_i} \quad (3.4)$$

The average (over the ten indices) was used to compare the general performance of the different datasets obtained after the application of OI and EnOI.

The Mean Absolute Error (MAE), defined by the following equation, was also considered:

$$MAE = \frac{1}{N} \sum_{i=1}^N |M_i - O_i| \quad (3.5)$$

It was used to quantify the improvements of the updated field (after applying OI or EnOI) with respect to the initial field (reanalysis for OI or ensemble mean of reanalysis for EnOI).

In order to verify the methods (OI and EnOI) a second validation was carried out based on a cross-validation strategy proposed by Panthou *et al.* (2012) defining two different subnetworks (*Near* and *Far*). Using this approach, both methods were applied again but now removing a group (validation group or subnetworks) from the initial set of stations. To create the *Near* subnetworks were used the following steps: 1) the inter-distances between the stations were estimated, 2) the pair of stations with the shorter inter-distance were selected, 3) the station from this pair with the minimum second shorter inter-distance to the remaining stations was removed from the calibration network and it was used as validation, 4) this process was repeated from the first step without using the station removed in the previous one, until the amount (e.g. percentage of the total) of desired validation points was reached. To create the *Far* subnetworks the first step mentioned above was used and for each station the shortest inter-distance was selected. The station with the maximum distance (between all the shortest inter-distances) was removed from the calibration network and it was used as validation, then, step 4 (mentioned above) was used.

Chapter 4

Results and discussion

The main results of the work are presented in the section 1.8 of the article (See Article 1). Complementary information is provided in this chapter.

After the application of both data assimilation methods (OI and EnOI), a comparison was first carried out at the stations that were used to update the initial fields. Figure 1.7 in the article compares the performance of OI and EnOI methods for each CPI. It can be seen that reanalysis performance depends on considered CPI and that EnOI datasets outperformed OI applied to each reanalysis independently. NRMSD values (averaged over all CPI) (see equation 1.13) for each year are presented in Figure 1.8 which confirms how the EnOI approach outperformed OI applied to each reanalysis for each year. Violin plots (Hintze & Nelson, 1998) were used to compare the density distribution of NE (see equation 3.4) for each dataset after the application of OI and EnOI; Figure 4.1). Box plots were also added to this figure. It is possible to see that, for each dataset, the median value is close to 0 (values close to 0 and small dispersion around it, represent better agreement with the observations). EnOI globally outperformed the other datasets with the smallest dispersion around 0. It can also be observed, as pointed out in the article (Figure 1.8) that EnOI performance is followed by JRA55, ERAI, MERRA and CFSR respectively. Cross-validation was then applied in order to assess the performance of OI and EnOI at points that were not used to upgrade the initial field (see more details in section 1.7). Figure 1.10 of the article compares MAE (see equation 1.14) for each index, configuration and datasets after applying the data assimilation methods compared with the initial fields. It can be seen that for each CPI and cross-validation network *Near* or *Far*

(and therefore even for points far away from the calibration network), estimated CPI fields were improved after applying OI or EnOI compared to the reanalysis values.

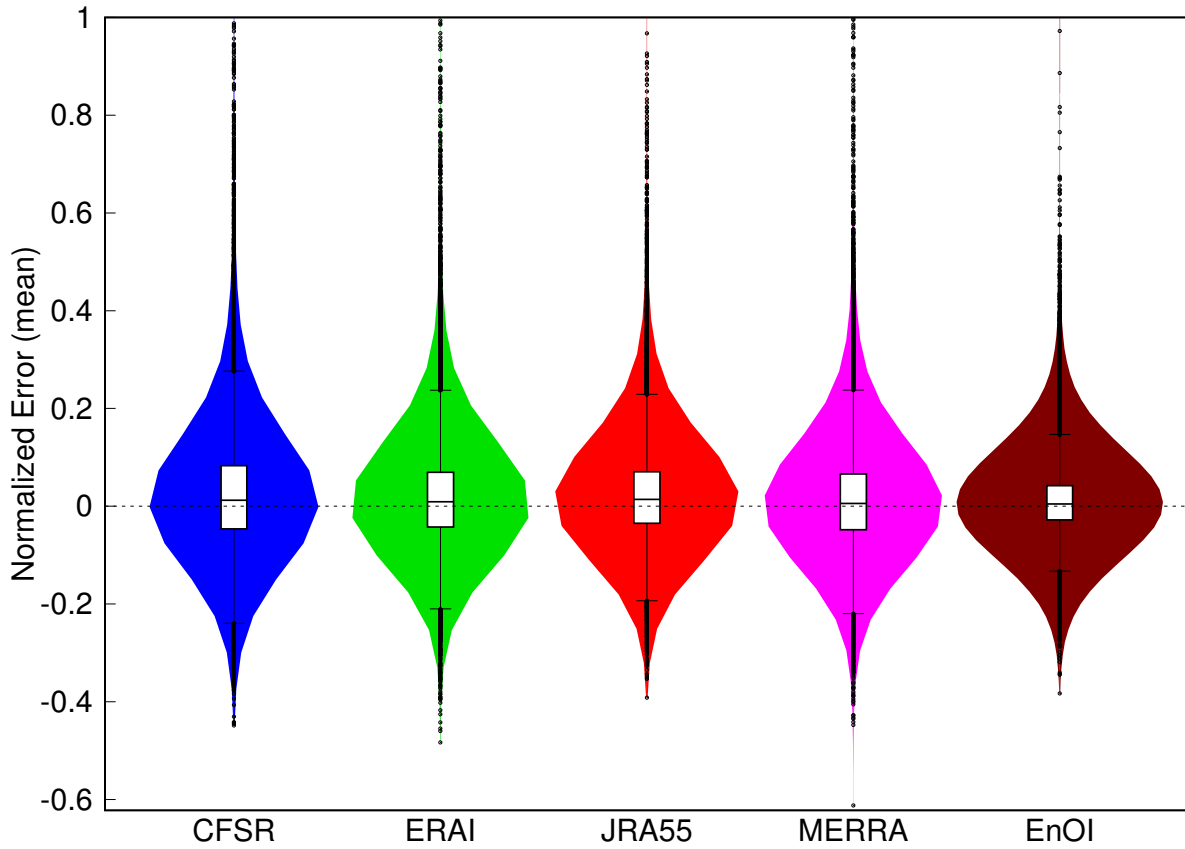


Figure 4.1 – Violin and box plots showing the NE (averaged over all ten indices) after the application of the data assimilation techniques. Boxes represent inter-quartiles range (first, Q1 and third, Q3), line inside the box the median value and outliers correspond to the black dots. Outliers are defined as values 1.5 times the interquartile distance ($IQR=Q3-Q1$) (whiskers) above or below Q3 or Q1 respectively.

In order to see the general improvements for each year over the study period, Figure 4.2 shows a heatmap of the annual values of V defined by:

$$V = \frac{1}{k} \sum_{i=1}^k \frac{MAE_{A(i)}}{MAE_{B(i)}} \quad (4.1)$$

where MAE_A is the Mean Absolute Error of the modified field (reanalysis after applying OI or EnOI), MAE_B the Mean Absolute Error of the initial field (reanalysis for OI or average of all reanalyses for EnOI) and k the number of CPI ($k = 10$). Each value of the heatmap (V) therefore corresponds to the mean V value over all CPI for *Near* or *Far* configurations with 10, 30, or 50% validation sites for a given year. It can be seen that for each year and each *Near* or *Far* network (and

therefore even for sites far away from the calibration network), estimated CPI fields were improved after applying OI or EnOI compared to the reanalysis values (all V values are smaller than one). Comparing the amplitudes of these improvements between the datasets for each subnetwork (*Near* or *Far* with 10, 30 or 50% validation sites) was carried out. It can be observed that CFSR is the reanalysis which displayed the largest improvements (compared with the initial field) over many years followed by MERRA. This figure complements the analysis presented in the article (Figure 1.10) where each CPI was analyzed independently over the study period and similar behaviours were observed (CPI fields were improved and the largest improvements were mostly concentrated between CFSR and MERRA).

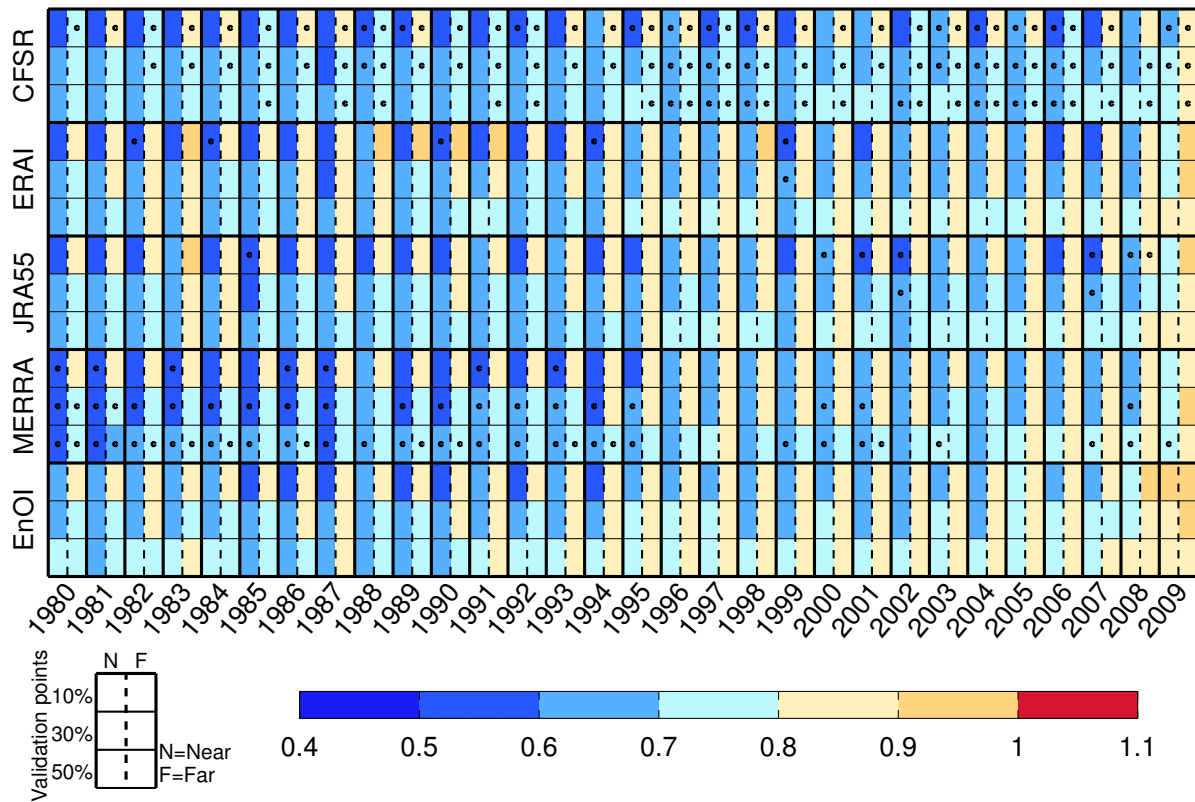


Figure 4.2 – Annual values of V (see equation 4.1) after applying OI to each reanalysis (top 12 rows) and after applying EnOI (bottom four rows) averaged over all CPI for the various cross-validation subnetworks (*Near* and *Far* configurations with 10, 30 and 50% validation sites). Black dots correspond to the datasets with the largest improvements for a given cross-validation subnetwork.

A comparison between each dataset was carried out. Figure 1.11 in the article summarizes the results of this comparison and shows that the ensemble approach outperforms the different reanalyses updated by OI in most cases. The dataset generated by the EnOI method was finally compared with the gridded dataset of NRCAN. Figures 1.12 and 1.13 compare both datasets through three

indices (PRmean, PR1mm and RX1day), showing that the CPI estimated from the EnOI dataset outperform NRCan. Figures 4.3 and 4.4 show the violin plots providing the density distribution of the differences or the errors (data minus observation) for each index.

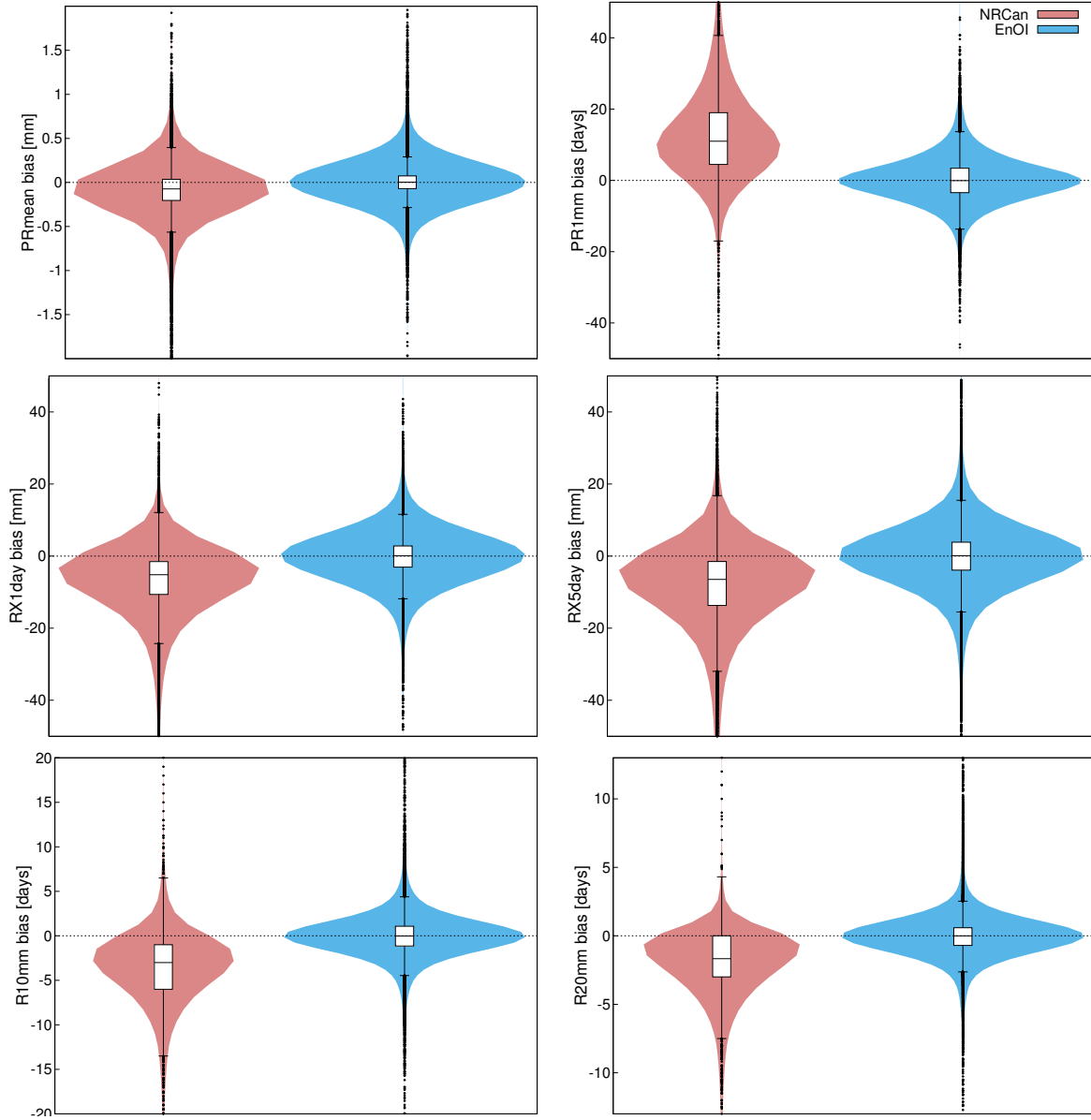


Figure 4.3 – Violin plots showing the bias between EnOI or NRCan and observed values at all the stations for PRmean, PR1mm, RX1day, RX5day, R10mm and R20mm. Boxes represent inter-quartiles range (first, Q1 and third, Q3), line inside the box the median value and outliers correspond to the black dots. Outliers are defined as values 1.5 times the interquartile distance ($IQR=Q3-Q1$) (whiskers) above or below Q3 or Q1 respectively.

It can be seen how the EnOI dataset outperforms NRCan for all indices (median values close to 0 and higher violin plot density around 0). NRCan dataset generally underestimates observed

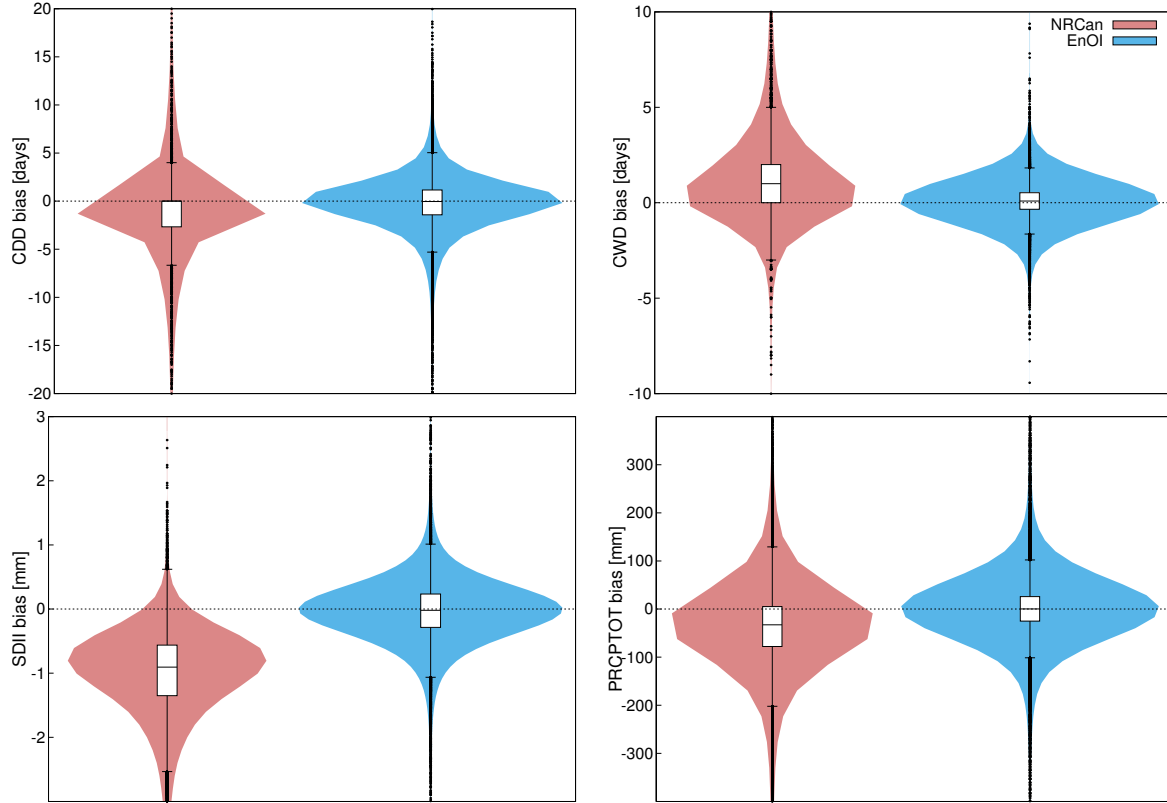


Figure 4.4 – Violin plots showing the bias between EnOI or NRCan and observed values at all the stations for CDD, CWD, SDII and PRCPTOT. Boxes represent inter-quartiles range (first, Q1 and third, Q3), line inside the box the median value and outliers correspond to the black dots. Outliers are defined as values 1.5 times the interquartile distance ($IQR=Q3-Q1$) (whiskers) above or below Q3 or Q1 respectively.

values for extreme indices such as RX1day, RX5day, R10mm, R20mm, SDII and PRCPTOT while it overestimates them for PR1mm and CWD. For CDD, median biases for NRCan are close to 0 but more dispersed than corresponding EnOI values. However, despite PRmean (from NRCan) shows underestimations compared to the observations, these remained small. These results are complementary to those presented in the article (section 1.8.3). Based on these results, it is recommended to use the NRCan dataset with caution for extreme precipitation estimation.

Chapter 5

Summary and Conclusions

Station network density in remote regions of Canada is small. It can be a huge challenge when defining a reliable reference climate datasets by using only this information. Reanalyses represent an interesting option that could be useful for those regions, as they provide consistent spatio-temporal gridded datasets. However, precipitation remains a problematic variable because it is a direct output of the numerical model and observations for this variable have not been explicitly used to create the reanalyses. Data assimilation techniques are interesting alternatives to achieve the goal of combining models outputs and observations in order to define a historical reference dataset.

The OI and EnOI methods were applied to improve the quality of the historical climate precipitation indices for Canada for a reference period of 30 years (1980-2009) where different reanalysis datasets (Climate Forecast System Reanalysis [CFSR], ERA-Interim [ERA-I], Japanese 55-year [JRA55] and Modern Era Reanalysis for Research and Applications [MERRA]) were combined with observational data (Daily precipitation records from the Adjusted and Homogenized Canadian Climate Data [AHCCD] developed by Environment Canada and from the stations operated by *Ministère du Développement Durable, de l'Environnement et de la Lutte contre les Changements Climatiques* [MDDELCC] from Quebec). Precipitation gridded dataset from Natural Resources Canada (NRCan) generated from stations records was also used, mainly for comparison purpose.

A group of ten CPI (defined by the Expert Team on Climate Change Detection and Indices [ETCCDI]) were analyzed, characterizing different features of precipitation or events that occur

at least once a year. Annual CPI series were considered in order to maximize the effective use of available observational datasets. Each reanalysis was then combined with the observational dataset using the OI approach while the EnOI method was applied combining the four reanalyses with the observational dataset.

A first analysis was performed aiming at comparing the OI and EnOI datasets at the same observation points and after by a cross-validation strategy. The estimated CPI show improvements compared to the initial fields, increasing also our ability to estimate them with some improvements at ungauged sites. The major improvements were obtained with the ensemble approach compared to the OI for each reanalysis independently. The climate precipitation indices obtained by the ensemble approach were regionally compared with the NRCan dataset, showing that NRCan should be used with caution, especially for extremes. The proposed dataset increases our ability to characterize the current climate in Canada and contributes to the development of climate change impact studies and adaptation strategies.

Part II

Article

Article 1

Defining a reference climate for Canada through the combination of observational and reanalysis datasets

Définition d'un climat de référence pour le Canada par la combinaison des ensembles de données d'observation et de réanalyse.

Authors

Alexis Pérez Bello, Alain Mailhot

Institut National de la Recherche Scientifique (INRS), Centre Eau Terre Environnement (ETE),
Québec city, Québec, Canada.

Author contribution statement

Prof. Alain Mailhot conceived, proposed and directed the presented work. Mr. Alexis Pérez Bello was responsible of implementation, adaptation and application of the methodology in a computational framework as well as analyzing and processing the different datasets to address the problematic presented. The results were analyzed and discussed by both authors. Mr. Alexis Pérez Bello was also responsible of writing the draft of the manuscript with important critical revisions and feedbacks from Prof. Alain Mailhot.

Ready for submission

Résumé traduit

La définition d'un climat de référence est un enjeu important pour le développement des futures stratégies d'impact et d'adaptation des changements climatiques. Cela reste encore un défi dans les régions où historiquement la couverture par des stations météorologiques est faible (p.ex. le Nord du Canada). Les réanalyses représentent une option intéressante pour réaliser cet objectif. Cependant, en considérant les biais et les inexactitudes des différentes réanalyses présentement disponibles, une validation et une correction de biais de ces ensembles de données sont encore nécessaires avant qu'ils puissent être utilisés. Dans cette étude, deux méthodes d'assimilation des données ont été considérées pour améliorer les champs climatiques de référence au Canada. Les méthodes d'interpolation optimale (OI pour Optimal Interpolation) et d'interpolation optimale d'ensemble (EnOI pour Ensemble Optimal Interpolation) ont été utilisées pour combiner quatre ensembles de données de réanalyses (CFSR, Era-Interim, JRA55 et MERRA) avec deux ensembles de données d'observation. Un total de 2160 stations météorologiques ayant au minimum 10 années valides d'enregistrements de précipitations et des données sur grille de Ressources Naturelles Canada ont été utilisées. Les valeurs annuelles de dix indices de précipitations climatiques (CPI pour Climate Precipitation Indices) couvrant une période de 30 ans (1980-2009) ont été estimées pour chaque ensemble de données et ont ensuite été combinées (réanalyse + observations) pour améliorer la représentation de chaque indice climatique au Canada. Une stratégie de validation croisée a finalement été appliquée et a montré que l'ensemble de données de référence développé améliore la capacité à estimer les indices climatiques aux endroits où aucune précipitation enregistrée n'est disponible. L'ensemble de données proposé a également été comparé à l'ensemble de données d'observation sur grille de Ressources Naturelles Canada (NRCan). Les résultats ont montré que l'ensemble de données NRCan devrait être utilisé avec prudence, en particulier pour les indices de précipitations extrêmes et dans les régions éloignées.

1.1 Abstract

Defining a reference climate is an important issue for climate change impact and the development of adaptation strategies. This remains a challenge in regions that were historically poorly covered by meteorological stations, such as Northern Canada. Reanalyses represent an interesting option to define a reference climate in such regions. However considering the bias and inaccuracies of the

various reanalyses actually available, some validation and bias-correction of these datasets are still needed before they could be used. In this paper, two data assimilation methods were considered to improve the reference climate fields over Canada. Optimal Interpolation (OI) and Ensemble Optimal interpolation (EnOI) methods were used to combine four reanalysis datasets (CFSR, Era-Interim, JRA55 and MERRA) with observations. A total of 2160 meteorological stations with minimally 10-year precipitation records were used. Annual values of ten Climate Precipitations Indices (CPI) covering a 30-year period (1980-2009) were estimated for each dataset and were then combined (reanalysis + observations) to improve the representation of each climate index across Canada. A cross-validation strategy was finally applied and showed that the developed reference datasets improve our ability to estimate climate indices at sites where no recorded precipitations are available. The proposed dataset (obtained from EnOI) was also compared to the gridded observational dataset from Natural Resources Canada (NRCan). Results showed that NRCan dataset should be used with caution especially for extreme CPI and in remote areas.

1.2 Introduction

In a context of climate change, characterizing historical climate is important to investigate how climate conditions may evolve in the future (Tebaldi *et al.*, 2006; Hatzaki *et al.*, 2010; Westra *et al.*, 2014). Historical climate can be characterized through various indices such as the Climate Precipitations Indices (CPI) proposed by the Expert Team on Climate Change Detection and Indices (ETCCDI) (Peterson *et al.*, 2001; Zhang *et al.*, 2011). These indices have been estimated directly from available observations around the world (e.g., Frich *et al.* (2002)) and also different gridded datasets of these indices have been produced by interpolating observational datasets (e.g., Alexander *et al.* (2006); Donat *et al.* (2013)). If this could be simply done in regions with dense meteorological network operating since many decades, it represents a real challenge in remote regions with short records and low density network as it is the case for many Canadian regions, especially in the northern parts (Mekis & Vincent, 2011).

Reanalyses represent an interesting alternative for such regions as they provide comprehensive and consistent spatio-temporal gridded datasets. Reanalyses consist of a background forecast model and data assimilation routine that combines available observational datasets with model forecast to generate uniform gridded datasets (Bosilovich *et al.*, 2008). They have been widely used in

different climate research applications (e.g., Alexeev *et al.* (2011); Zou *et al.* (2014); Lindsay *et al.* (2014)). Extreme indices for precipitation and temperature from reanalyses have also been used as reference for evaluation of climate models (e.g., Sillmann *et al.* (2013)). Their consistency with global gridded observational datasets have been verified (Donat *et al.*, 2014), showing improvements in representation of the spatial patterns after 1979 when satellite data were included in the assimilation process. Despite their overall good spatio-temporal representation of many variables, some of them (e.g. precipitation) are still not well represented in some regions, mainly because observational datasets for these variables have not been explicitly used in the assimilation process (Saha *et al.*, 2010; Dee *et al.*, 2011; Kobayashi *et al.*, 2015; Rienecker *et al.*, 2011). Evaluating daily climate indices over the northern parts of Canada, Diaconescu *et al.* (2017) recommend caution when using extreme daily precipitation indices from reanalyses as reference datasets for these regions. Reanalyses therefore need to be bias corrected before they could be used. Their performances depend on the variable considered, the data assimilation routine, the assimilated observational datasets, and the forecast model (Lorenz & Kunstmann, 2012).

One possible option for improving the CPI representation is through the combination of reanalysis datasets with *in situ* observations through data assimilation methods. These methods combine model fields with observational datasets to create an improved model dataset (Kalnay, 2002; Bertino *et al.*, 2007). They can be incorporated into the modeling process through two different forms: online mode, which increases the quality of the initial conditions sequentially for new model simulations and off-line mode, which is used to create the best estimation, combining the final outputs of the model with observational data (Candiani *et al.*, 2013; Matsikaris *et al.*, 2015).

In this work the off-line mode was applied using the Optimal Interpolation (OI) and Ensemble Optimal Interpolation (EnOI) methods. OI was applied to each reanalysis independently and EnOI to all reanalyses combined. Both methods are quite similar, their main difference being in how the background error covariance is defined (Ren & Hartnett, 2017). These methods were applied because they are relatively straightforward, computational inexpensive and they take into account the uncertainty of the different datasets used. OI method has been used to merge different sources of precipitations in some regions of the world (e.g., Häggmark *et al.* (2000); Mahfouf *et al.* (2007); Soci *et al.* (2016)). Furthermore, significant improvements were obtained with its application at global scale by Nie *et al.* (2016) merging satellite information with observations and model predictions.

This method has been also used for daily retrospective estimation of streamflow in ungauged rivers by Lachance-Cloutier *et al.* (2017).

The objective of the paper is to improve the historical CPI datasets for Canada combining observations and reanalyses through the OI and EnOI methods. It is organized as follows. Section 1.3 describes the observational datasets and reanalyses considered in this study. The CPI, data assimilation techniques (OI and EnOI), and the cross-validation procedure are briefly described in Sections 1.4, 1.5, 1.6 and 1.7 respectively. Section 1.8 presents the results followed by a summary and a discussion (Section 1.9).

1.3 Datasets

Five gridded datasets were considered in this study (Table 1.1). The first four ones are widely used reanalyses while the fifth one is the interpolated precipitation dataset from Natural Resources Canada (NRCan) generated from stations records using a thin plate spline smoothing algorithm (ANUSPLIN) (Hutchinson *et al.*, 2009; McKenney *et al.*, 2011). This dataset was used for comparison purpose and to estimate missing values in station records (further details are provided in Section 1.5), also it provides information with good spatial resolution (10 km spatial resolution) which covers Canada (the study domain) and covers a relatively long period (1950 to 2012). Previous studies show underestimation for precipitation values in this dataset (e.g. Hutchinson *et al.* (2009); Bajamgigni Gbambie *et al.* (2017); Diaconescu *et al.* (2017)). Despite its weaknesses, to our knowledge, there is no other dataset having these characteristics (accessibility, domain size, spatio-temporal resolution and period covered) that can be used for the study area considered in this work. Some other datasets exist but they cover short periods (e.g. CaPA) or cover only Quebec (e.g. MDDELCC dataset; see Bajamgigni Gbambie *et al.* (2017)). Since each dataset covers a different period, the common period, 1980-2009, was considered in the following.

Daily precipitation records from the Adjusted and Homogenized Canadian Climate Data (AHCCD) developed by Environment Canada (Mekis & Vincent, 2011) and from the stations operated by *Ministère du Développement Durable, de l'Environnement et de la Lutte contre les Changements Climatiques* (MDDELCC) from Quebec were also used. Only stations with at least 10 valid years during the 30-year period 1980-2009 were retained. A valid year is defined as a year

Table 1.1 – Gridded datasets used in this study (acronyms appearing in parenthesis are used to refer to these datasets).

Gridded datasets	Period	Spatial resolution	Temporal resolution	Reference
Climate Forecast System Reanalysis (CFSR)	1979-2009	$0.312^\circ \times 0.312^\circ$	Hourly	Saha <i>et al.</i> (2010)
ERA-Interim (ERA-I)	1979-2012	$0.75^\circ \times 0.75^\circ$	12 hours	Dee <i>et al.</i> (2011)
Japanese 55-year Reanalysis (JRA55)	1958-2013	$0.56^\circ \times 0.56^\circ$	3 hours	Kobayashi <i>et al.</i> (2015)
Modern-Era Retrospective Analysis for Research and Applications (MERRA)	1979-2012	$0.5^\circ \times 0.66^\circ$	Hourly	Rienecker <i>et al.</i> (2011)
Precipitation data from Natural Resources Canada (NRCan)	1950-2012	$0.083^\circ \times 0.083^\circ$	Daily	McKenney <i>et al.</i> (2011)

with less than 20 % missing daily values (Panthou *et al.*, 2014), for a total of 2160 stations across Canada. Figure 1.1 presents the map of these stations while Figure 1.2 presents the histogram of the number and percentage of stations with given numbers of valid years during the 1980-2009 period. Figure 1.1 clearly shows that stations are unevenly distributed and mainly concentrated in the southern parts of the country. Figure 1.2 confirms that available records contain many missing data (only 10.8 % of the stations have records covering the complete 1980-2009 period).

It is important to note that not all the stations considered in this study were used to create the NRCan dataset (for instance, some stations operated by the MDDELCC in Quebec were not used). Also it is well known that interpolated precipitation, especially extreme values, at daily time scale can be highly uncertain in regions with sparse network coverage (Gervais *et al.*, 2014).

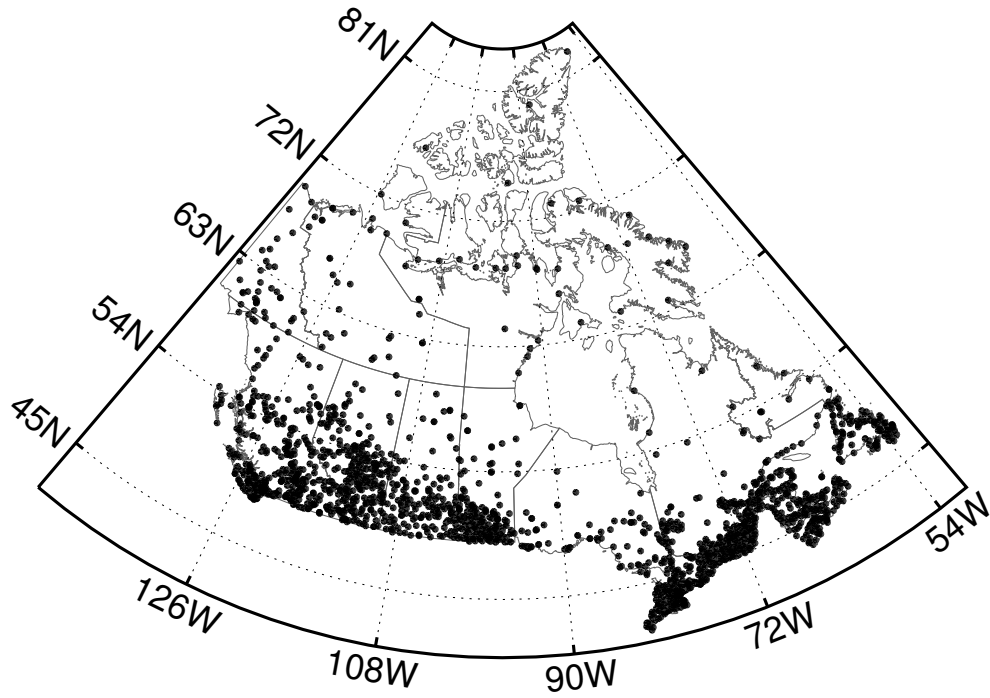


Figure 1.1 – Map of the stations considered in this study.

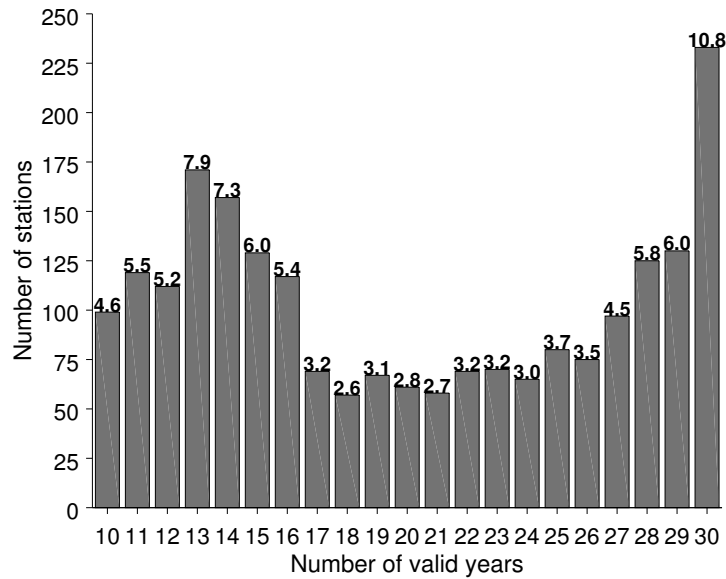


Figure 1.2 – Number and percent (above each bar) of stations with given numbers of valid years.

1.4 Climate Precipitation Indices and Super-Observations

Different Climate Precipitation Indices (CPI) defined by the Expert Team on Climate Change Detection and Indices (ETCCDI) as well as mean annual precipitation (PRmean) have been

considered in this study (Table 1.2). Annual CPI values were first estimated as well as climatological mean over the 1980-2009 period. CPI were estimated at each native grid point of each dataset (reanalysis or NRCan dataset). A second-order conservative remapping (Jones, 1999) was then applied to each CPI gridded dataset created from ERAI, JRA55 and MERRA using the CFSR grid as reference grid. OI and EnOI methods were applied once CPI gridded datasets from each reanalysis were put on a common grid.

Table 1.2 – Climate Precipitation Indices, adapted from Zhang *et al.* (2011)

Index	Description	Units
PRmean	Annual mean of daily precipitation.	mm
PR1mm	Number of wet days (daily precipitation ≥ 1 mm/day).	days
RX1day	Annual maximum 1-day precipitation amount.	mm
RX5day	Annual maximum 5-day precipitation amount.	mm
SDII	Simple daily intensity index (The ratio of annual total precipitation to the number of wet days ≥ 1 mm).	mm
R10mm	Annual number of days when precipitation ≥ 10 mm.	days
R20mm	Annual number of days when precipitation ≥ 20 mm.	days
CDD	Annual maximum number of consecutive days when precipitation is below 1 mm.	days
CWD	Annual maximum number of consecutive days when precipitation ≥ 1 mm.	days
PRCPTOT	Annual total precipitation from days ≥ 1 mm.	mm

As shown in Figure 1.1, station density in southern Canada can be high. Observations from neighboring stations can therefore be highly correlated at the daily time scale. Furthermore reanalysis do not properly represent small scale processes (Vihma *et al.*, 2014). Super-observations (Daley, 1991; Oke *et al.*, 2006), also called *upscaled observations* (Cherubini *et al.*, 2002), were therefore created using the reanalysis grid with the highest resolution as reference grid (i.e. CFSR).

Super-observations values were estimated by combining station values within the same grid cell using the following expression:

$$\bar{x} = \frac{\sum_{i=1}^n V_i X_i}{\sum_{i=1}^n V_i} \quad (1.1)$$

where n is the number of stations inside a given grid cell, V_i is the number of valid days per station for each year, while X_i represent annual CPI value for year i . Longitude and latitude of the super-observations were also estimated using a similar expression.

Figure 1.3 shows the number of stations and corresponding super-observation with valid records for each year of the 1980-2009 period. It shows that, after a peak in the 1988-1992, the number of stations quickly decreased to reach a minimum value in 2009. The spatial distribution of the super-observations changes from year to year depending of the number and location of stations available each year. These super-observations networks were finally used to apply OI and EnOI for each year independently.

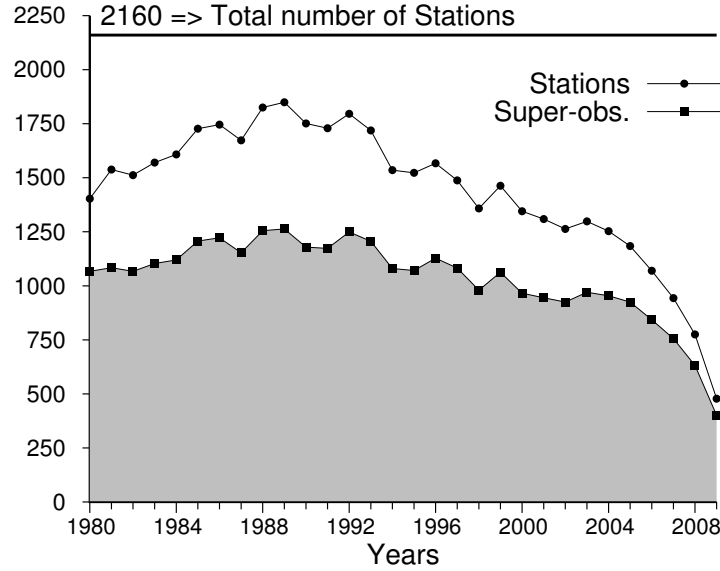


Figure 1.3 – Total number of stations and super-observations available each year.

1.5 Optimal Interpolation (OI)

This section provides a brief description of the Optimal Interpolation (OI) method and how it was applied in this paper. A more general and comprehensive description can be found in Daley (1991); Kalnay (2002); Sen (2009).

Optimal interpolation, also known as *statistical interpolation*, is a widely used method (e.g., Mahfouf *et al.* (2007); Lespinas *et al.* (2015)) to create an analysis field by combining a background state, or model forecast, with the observation network (Daley, 1991). The analysis value $\psi_A^{(j)}$ at grid point j is estimated by the background field $\psi_B^{(j)}$ plus the linear combination of the differences between the observations $\psi_O^{(i)}$ ($i = 1 \dots n$) where n is the number of observations close to this grid point and the background values at the observation site $\psi_B^{(i)}$. The so-called analysis equation is defined as:

$$\psi_A^{(j)} = \psi_B^{(j)} + \sum_{i=1}^n W^{(j,i)} (\psi_O^{(i)} - \psi_B^{(i)}) \quad (1.2)$$

Assuming that the background error only depends on distance (so-called homogeneous conditions), is independent of the direction (isotropic), and that the observation errors are uncorrelated, then, W , the interpolation weights matrix, can be estimated by the following equation Daley (1991):

$$[B_{EC} + \mu I]W = b_{EC} \quad (1.3)$$

in which B_{EC} is the background error correlation matrix between the station values, b_{EC} is the background error correlation vector between the observations and the grid point to analyze from the background field, I is the identity matrix and $\mu = O_{EV}/B_{EV}$ where O_{EV} and B_{EV} are the observation and background error variance respectively. Parameter μ can be adjusted such that the weights granted to observations can be increased or reduced (Kalnay, 2002). O_{EV} and B_{EV} for each index and reanalysis respectively were estimated according to the methodology described in Section 4.3 of Daley (1991).

The available CPI annual series over the studied 30-year period were used to estimate the spatial dependence of the correlation function $B_{EC}(x)$ in equation (1.3). The values at each observation point from the background field were obtained for each year by mean of bilinear interpolation and the missing values in the observational dataset were filled out with the CPI calculated from NRCan

dataset. Then, correlations R_{ij} of the differences between the observations and background field at site i and j were calculated using the following equation (Sen, 2009; Daley, 1991):

$$R_{ij} = \frac{\overline{(O_i - B_i)(O_j - B_j)}}{\sqrt{\overline{(O_i - B_i)^2} \overline{(O_j - B_j)^2}}} \quad (1.4)$$

where the overlined variable corresponds to average of annual values over the reference period. The mean correlations over each 30-km distance intervals were then computed. Many different spatial correlation functions $R(x)$ were used and the following expression was finally selected and adjusted to the estimated values (Sen, 2009):

$$R(x) = a + be^{(-x/c)} \quad (1.5)$$

where x is the distance between stations and a , b and c are fitting parameters. Figure 1.4 shows an example of the mean correlation estimated for each distance interval with the corresponding fit of equation 1.5 for the RX1day index estimated from CFSR.

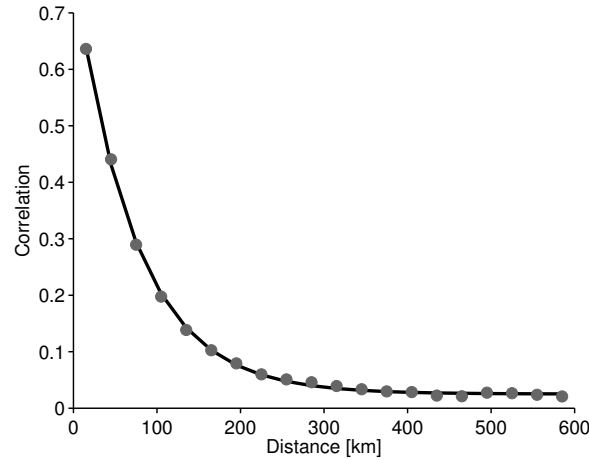


Figure 1.4 – Estimated correlation function (continuous curve) with corresponding discrete values for the RX1day from CFSR.

Due to the averaging process over successive intervals, Figure 1.4 shows that zero distance correlation has to be extrapolated. Using equation 1.5 and setting $x = 0$, the isotropic background error correlation B_{EC} used in equation 1.3 is given by (Sen, 2009):

$$B_{EC}(x) = \frac{R(x)}{R(0)} \quad \text{or} \quad B_{EC}(x) = \frac{a + be^{(-x/c)}}{a + b} \quad (1.6)$$

The relative expected analysis error ε for each grid point can be estimated by the following equation (Kalnay, 2002; Sen, 2009; Daley, 1991):

$$\varepsilon = 1 - \sum_{i=1}^n W^{(i)} b_{EC}^{(i)} \quad (1.7)$$

where values of ε close to 0 means a biggest influence or weight of neighbouring observations in the estimation of the CPI value at this specific grid point while values of ε close to 1 means a biggest contribution of the background field. W and b_{EC} were explained above when equation 1.3 was presented.

1.6 Ensemble Optimal Interpolation (EnOI)

This section provides a brief description of Ensemble Optimal Interpolation (EnOI) method and how it was applied in this paper. A more general and comprehensive description can be found in Evensen (2003).

The EnOI is an approximation of the Ensemble Kalman filter (EnKF) in which an ensemble of model states in long time integration (stationary) is used to compute the analysis in the space spanned and only one model state is used as a background field (Oke *et al.*, 2006). Defining a matrix E with the ensemble members:

$$E = [M_1, M_2, ..., M_n] \quad (1.8)$$

where n is the number of ensemble members which are stored (vector form) in each column of E . For each CPI, the ensemble members are the 30 years of each reanalysis combined ($n = 120$ members). The anomaly or ensemble perturbation matrix E_n can be defined by:

$$E_n = E - E_m \quad (1.9)$$

where E_m is the ensemble mean matrix in which each column has the same ensemble mean vector calculated from all the ensemble members stored in E . Then, the ensemble covariance matrix E_{CM} can be estimated by:

$$E_{CM} = \frac{\alpha E_n E_n^T}{n - 1} \quad (1.10)$$

where the superscript T means a matrix transpose and the parameter $\alpha \in (0,1]$ scaled the magnitude of the model anomalies. This parameter is introduced to reduce the variance produced by the ensemble over a long time period that can overestimate the instantaneous background field error variance (Fu *et al.*, 2011; Evensen, 2003). In our case, the value $\alpha = 1$ was used.

The EnOI analysis field was estimated using analysis equation 1.2 described in section 1.5 and the following weight matrix W (Oke *et al.*, 2006):

$$W = C_F \circ E_n \alpha (H E_n)^T \left[C_F \circ H E_n \alpha (H E_n)^T + (n - 1) R \right]^{-1} \quad (1.11)$$

in which H is an operator that interpolate from model grid points to observation sites (bilinear interpolation), R is the observation error covariance matrix, C_F is the correlation function matrix used to implement the localisation (each analysis grid point is calculated using only the nearest stations, (Evensen, 2003)), and “ $\circ$ ” represents the Hadamard product, that basically multiplies element by element two matrices of the same dimensions (Oke *et al.*, 2006).

The background field used in the EnOI was the ensemble mean over the four reanalysis for each year. The observation errors were assumed uncorrelated as in section 1.5 and therefore the observation error covariance matrix R was diagonal with values equal to the observation error variance O_{EV} . The correlation function was the same as equation 1.6 and the parameters a , b , c for this equation were the mean parameters over the four reanalyses. The ensemble O_{EV} was estimated using the following equation:

$$O_{EV} = \frac{1}{k} \sum_{i=1}^k \frac{X_i}{k} \quad (1.12)$$

where k is the number of reanalyses (in our case 4) and X_i is the O_{EV} for each reanalysis. The O_{EV} differs from one reanalysis to another because it was estimated from the relationship between the observations and the background values at the observation sites (see section 1.5 and Section 4.3 of (Daley, 1991)).

1.7 Cross-validation procedure

In order to verify the methods (OI and EnOI) and check how the quality of CPI was improved in sites without observations, both methods were applied again but now removing a group (validation

group) from the initial set of stations. To address the selection problem of this validation group, the subnetwork cross-validation approach proposed by Panthou *et al.* (2012) was applied. Validation networks are generated by eliminating stations at varying distances in order to control the average distance between sites of the calibration and validation networks. Two types of calibration/validation networks were defined. *Far* networks were designed to assess the performance of the developed datasets to estimate values at distant locations from the calibration points, while *Near* networks were designed to assess the performance for locations close to the calibration points (Panthou *et al.*, 2012). Different validation networks including 10, 30, and 50 % of the stations from the initial network were defined for the *Far* and *Near* network configurations.

Figure 1.5 shows an example of *Near* (left panel) and *Far* (right panel) sub-networks including 10% validation sites (blue) and their corresponding 90% calibration sites (red). As explained in Section 1.4, the number of super-observations may change from year to year and therefore calibration and validation networks may change over time as well. Figure 1.6 presents violin plots (Hintze & Nelson, 1998) and boxplots of the distributions of distances between all validation sites with their closest calibration sites for *Near* and *Far* subnetworks respectively over the study period. For 10% validation points, the *Near* subnetwork corresponds to the configuration with the smallest distances between calibration and validation sites while the *Far* configuration has the largest distances. These distances increase as the fraction of validation sites increases for the *Near* subnetworks, while they decrease for the *Far* subnetworks. Furthermore, these distances decrease rapidly for *Far* configuration as the fraction of validation sites increases which is consistent with the spatial distribution of stations showed in Figure 1.1, where the network density is much higher in the southern part of Canada and only a small fraction of stations covers the Northern regions.

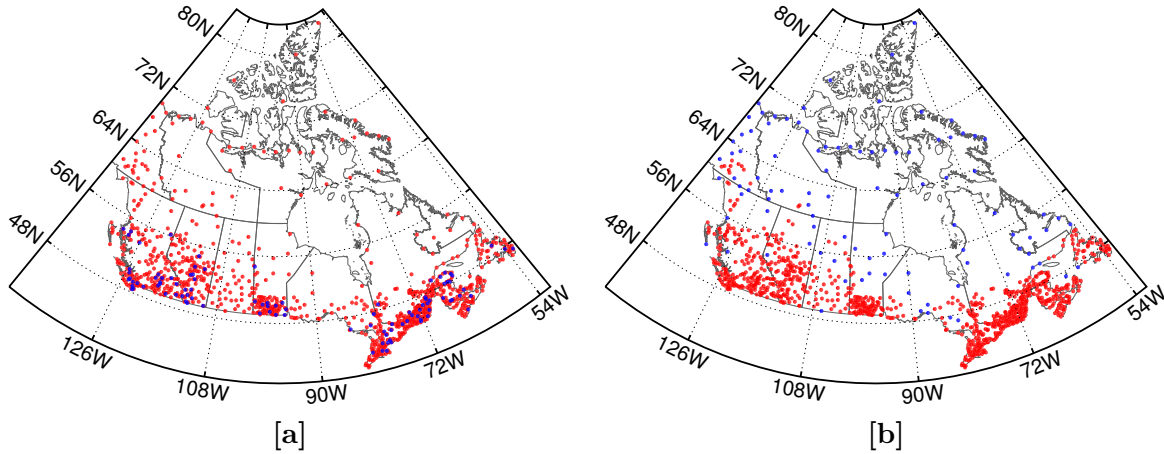


Figure 1.5 – Spatial distribution of calibration (red) and validation (blue) sites on year 1980 for (a) *Near* and (b) *Far* subnetworks with 10% validation stations.

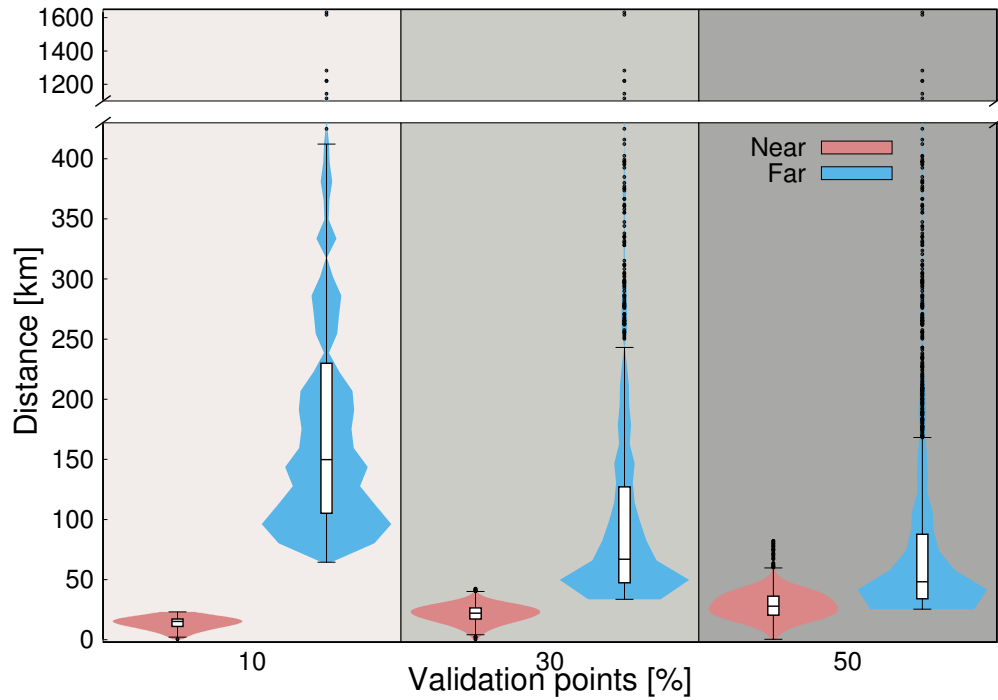


Figure 1.6 – Violin and box plots showing the distances between all validation sites and their closest calibration sites for *Near* and *Far* subnetworks respectively over the study period. Boxes represent inter-quartiles range (first quartile $Q1$ and third quartile $Q3$), line inside the box the median value and outliers correspond to the black dots. Outliers are defined as values 1.5 times the interquartile distance ($IQR=Q3-Q1$) (whiskers) above or below $Q3$ or $Q1$ respectively.

1.8 Results and Discussion

1.8.1 Comparison of reanalysis, OI, and EnOI datasets

Taylor diagrams (Taylor, 2001) were used to compare CPI annual series at stations with: a) reanalysis CPI values; b) reanalysis CPI values after the application of OI; c) reanalysis CPI values after the application of EnOI (Figure 1.7). The statistics displayed in the Taylor diagrams (Figure 1.7) were calculated for each CPI by combining the annual CPI values from all stations.

Reanalysis performance depends on considered CPI, the largest Normalized Root Mean Square Difference (NRMSD) being observed for the CFSR CWD index and the smallest for the JRA55 PRmean, R10mm and PRCPTOT indices. The reanalysis with the best performance is JRA55 while CFSR and MERRA displayed the poorest performance. Correlation coefficients for all indices are larger than 0.5. After applying OI, most indices display correlation values larger than 0.9, NRMSD smaller than 0.5 and normalized standard deviation (NSD) close to 1. Combining the various reanalyses through EnOI (Figure 1.7c) still improved the overall performance as the NRMSD for all indices is smaller than 0.3 and correlation coefficient larger than 0.95. The indices with the largest NRMSD are CDD, RX1day, and CWD. These results are not really surprising since performance is assessed at the same sites used for the data-assimilation process. They however confirm that applying OI or EnOI improves the estimates at station sites.

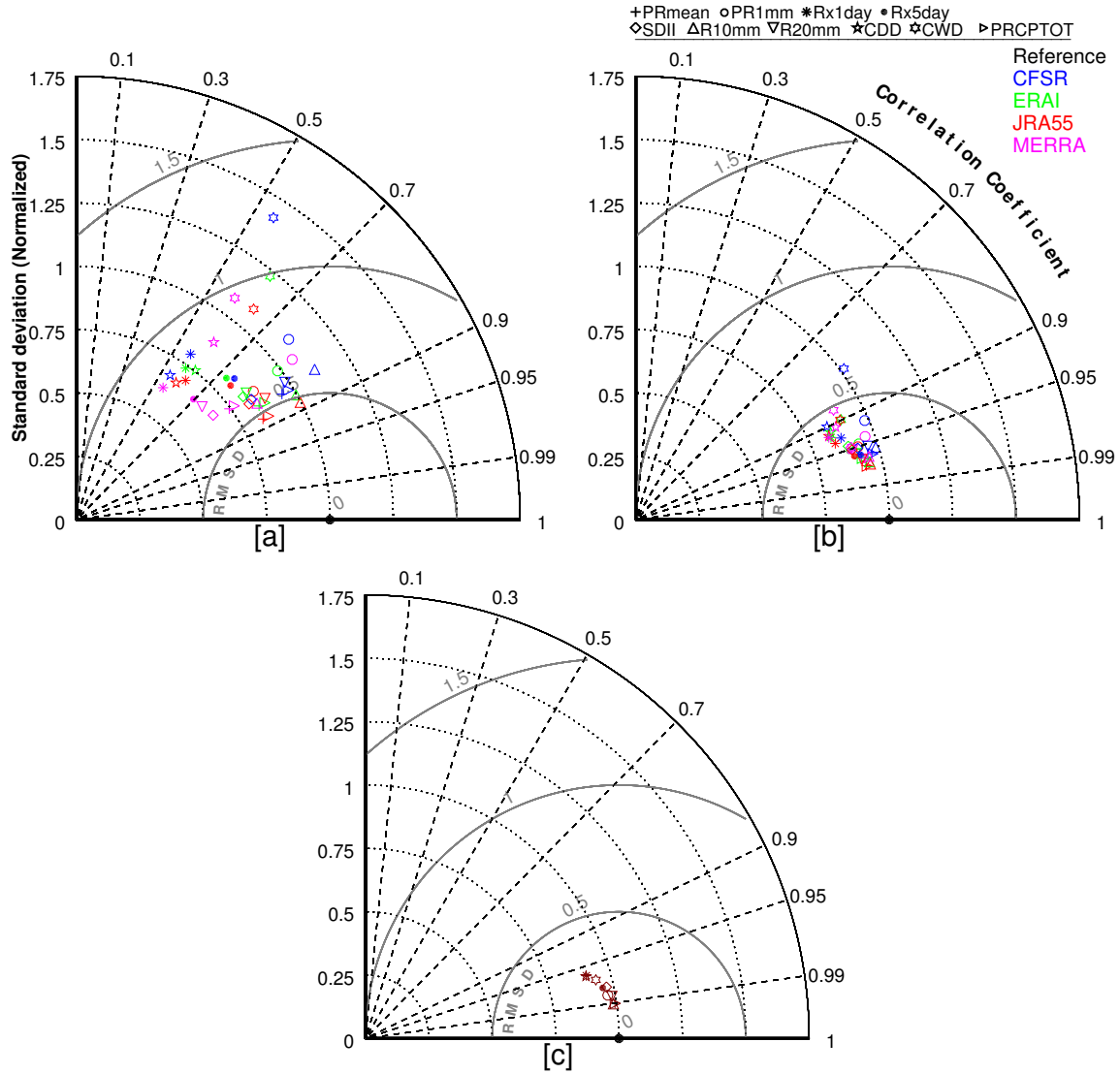


Figure 1.7 – Taylor diagrams comparing the observed and estimated annual CPI values over the 1980-2010 period at all station sites: a) before applying OI; b) after applying OI; c) after applying EnOI.

Figure 1.8 shows the annual series of NRMSD (averaged over all indices) as estimated by the following equation:

$$NRMSD_i = \frac{1}{k} \sum_{j=1}^k SNRMSD_{i,j} \quad (1.13)$$

where k represents the number of CPI ($k = 10$) and $SNRMSD_{i,j}$ represents the Spatial NRMSD calculated over all stations for a specific year i ($i = 1980, \dots, 2009$) and a specific CPI j . Figure 1.8 shows that EnOI datasets outperformed OI applied to each reanalysis for each year. These results confirm the effectiveness of the EnOI approach compared to the OI independently applied to each reanalysis.

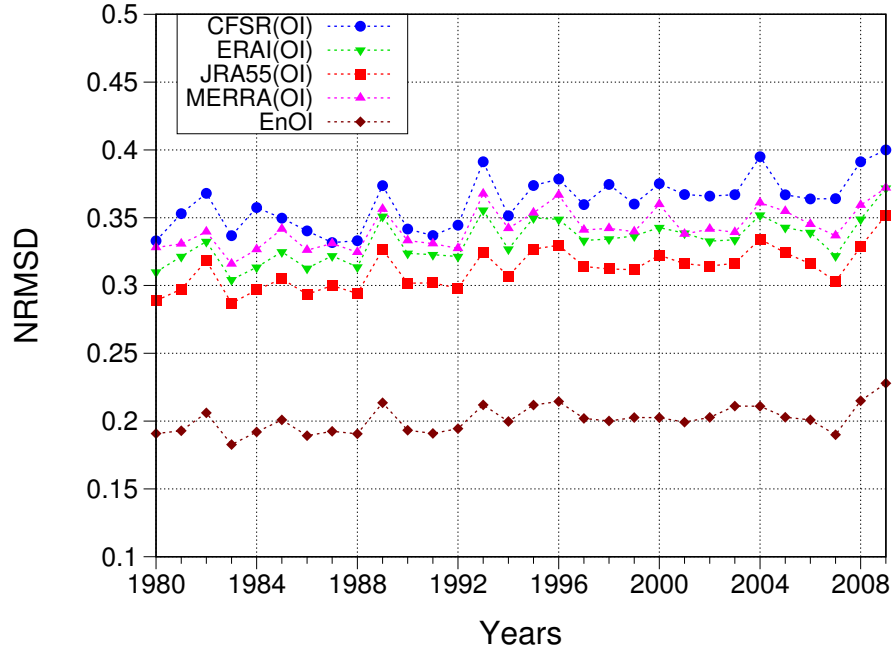


Figure 1.8 – Annual NRMSD (averaged over all stations and all CPI) values normalized by the observational standard deviation (σ over all valid stations for a given year) for each reanalysis after applying OI and EnOI.

1.8.2 Cross-validation

The distance between each land-based grid point over the study area and the closest station is presented in the Figure 1.9a. All stations over the 30-year study period were considered despite that their number and spatial distribution change from year to year due to missing data. In fact no year of the 1980-2009 period includes all stations as shown in Figure 1.3. Therefore, Figure 1.9a shows the areas where the estimated CPI were more influenced by the nearby station values (areas in blue tones) or by the initial field (reanalysis)(areas in red tones) values after the application of OI and EnOI. Figure 1.9b provides the relative expected average CPI error values, estimated using equation 1.7, as a function of distances to the closest station.

Despite CPI values at each grid point are estimated using neighboring stations, and not only the closest station, figure 1.9b provides valuable information about the relationship between expected CPI error and distance to the station having the largest impact on estimated values. It is possible to see that the contribution from observations to the analysis field decreases more quickly with distance for some indices (e.g., RX5day, RX1day, R20mm, CDD and CWD). Indices with the largest contribution from neighboring observations for a given distance are PRCPTOT, PRmean

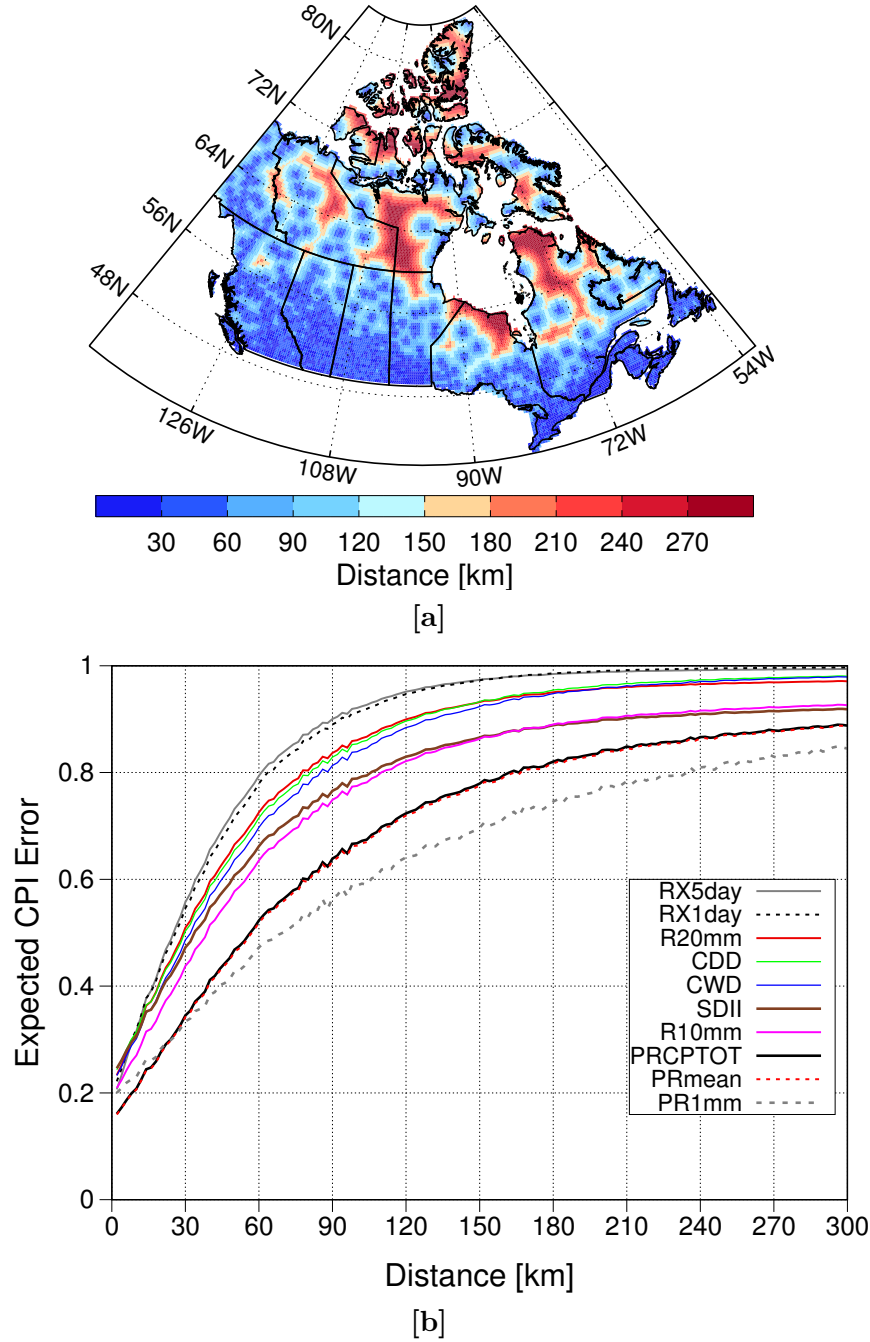


Figure 1.9 – (a) Distance between each grid point and the closest station and (b) expected relative CPI error (equation 1.7) as a function of the distance to closest station.

and PR1mm. It can be seen that for some remote regions the main contribution is from the reanalysis field and the overall quality of the analysis field highly relies on the quality of the reanalysis field.

Cross-validation was then applied to assess the performance of OI and EnOI at points that were not used to upgrade the initial field. Figure 1.10 shows a heatmap with the values of V_1 defined by:

$$V_1 = \frac{MAE_A}{MAE_B} \quad (1.14)$$

where MAE_A is the Mean Absolute Error of the annual CPI values at validation sites of the analysis field (reanalysis after applying OI or EnOI) and MAE_B the Mean Absolute Error of the annual CPI values of the initial or background field (each reanalysis value or average reanalysis value). It can be seen that for each CPI and cross-validation network *Near* or *Far* (and therefore even for points far away from the calibration network), estimated CPI fields were improved after applying OI or EnOI compared to reanalysis values (all V_1 values are smaller than one). It is possible to see also that the largest V_1 values (smallest improvement compared to background field) are, in general, for the *Far* configuration 10%, which represents the configuration with the largest averaged distance between calibration and validation networks (see Figure 1.5b and 1.6). When the percentage of validation points increases to 30 or 50 %, V_1 decreases as validation points are getting closer to calibration points (see Figure 1.6). Reanalyses cannot be compared to each other in Figure 1.10 since reference datasets are different in each case (denominator of eq. 1.14). For instance, one dataset can display a large improvement but the initial performance could have been poor while another one can display a small improvement but have a good initial performance. The V_1 values for the various datasets were compared for each subnetwork (*Near* or *Far* with 10, 30 or 50% validation points) and each CPI. Black dots on Figure 1.10 identify the dataset with the largest improvement (smallest V_1) for each subnetwork and each CPI. CFSR and MERRA were the reanalysis with the largest improvements for most subnetworks. The CPI with the smallest improvements, especially for *Far* subnetworks, were RX1day and RX5day while PR1mm stands out as the one with the largest improvements. These results are consistent with Figure 1.9b where contributions from observations to RX1day and RX5day decreased as distance to closest station increases while PR1mm is the CPI with the largest contribution from observation for closest station, more than 30 km from the considered point.

A similar heatmap is presented in Figure 1.11 where the values of V_2 are defined by:

$$V_2 = \frac{RMSD_A}{\sigma_o} \quad (1.15)$$

where $RMSD_A$ is the Root Mean Square Difference of the analysis field (reanalysis after applying OI or EnOI) and σ_o is the standard deviation for each CPI of annual values at all validation sites. In this case each dataset is compared to the same reference dataset and therefore it is possible to

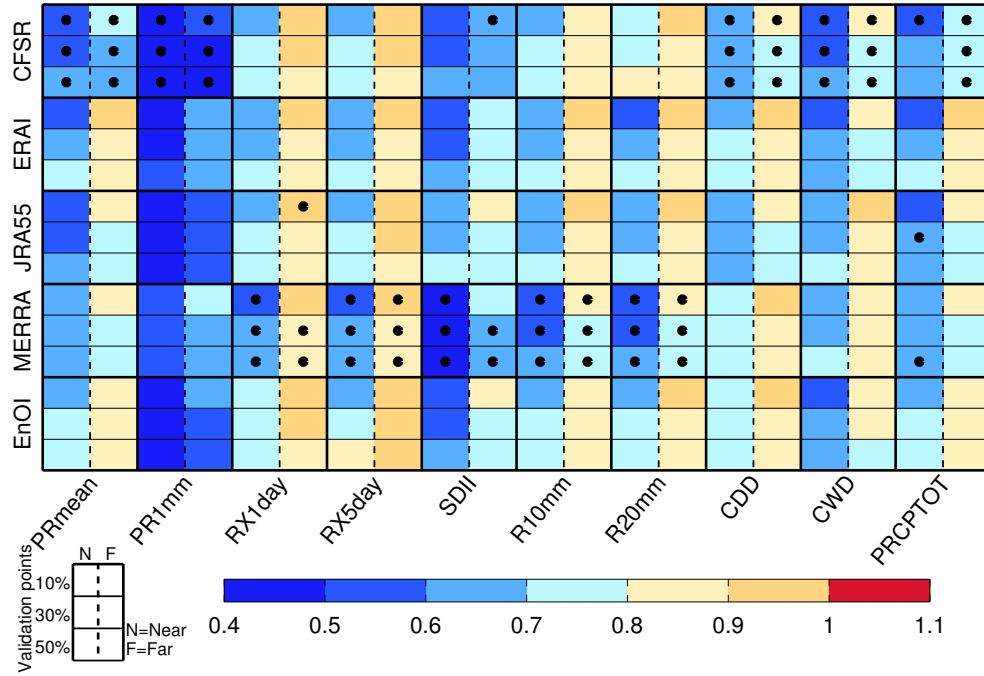


Figure 1.10 – Mean Absolute Error of the analysis field (MAE_A) normalized by the Mean Absolute Error of the background field (MAE_B) over all validation sites for the entire period (1980-2009) (equation 1.14) and each CPI (x-axis). *Near* (N) and *Far* (F) configurations with 10, 30, and 50% validation points after applying OI to the four reanalyses (top four rows) and after applying EnOI (bottom row) are presented. Black dots identify, for each CPI, the datasets with the lowest V_1 values for a given subnetwork (N or F with 10, 30, or 50% of validation points).

select the datasets displaying the best performance for each CPI. Figure 1.11 shows that, for most CPI and subnetworks, EnOI outperformed the OI datasets (36 CPI/subnetworks combinations, representing 60% of all CPI/subnetwork combinations), followed by the analysis field based on OI applied to JRA55 (21 CPI/subnetwork combination for 35% of all CPI/subnetwork combinations) and finally by the analysis field based on OI applied to MERRA and CFSR (3 CPI/subnetworks combinations for 5% of all cases). It can be seen also that EnOI outperformed the other datasets for all indices for the *Far* configuration with 10% validation points corresponding to the case where the average distance between calibration and validation sites is the largest (Figure 1.5b and 1.6).

These results show that applying OI and EnOI methods led to an improved representation of the fields of CPI across Canada compared to the initial field (single reanalysis) for all CPI and even at ungauged sites far from the sites that were used for OI and EnOI calibration. Also it shows that the EnOI in most cases outperformed OI. This last result is interesting as it shows that combining all the available information (all reanalyses) through EnOI is the best strategy.

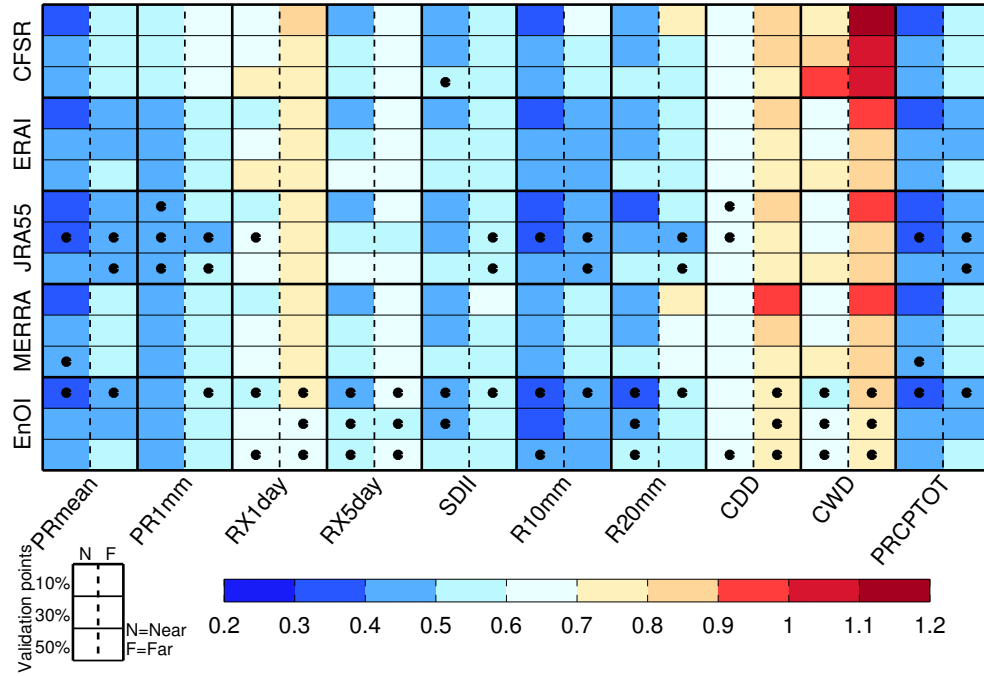


Figure 1.11 – RMSD of the analysis field ($RMSD_A$) normalized by the standard deviation for each CPI (σ_B) of annual values at all validation sites for the entire period (1980-2009) (equation 1.15) and each CPI (x-axis). *Near* (N) and *Far* (F) configurations with 10, 30 and 50% of validation points after applying OI to the four reanalyses (top four rows) and after applying EnOI (bottom row) are presented. Black dots identify the datasets with the best performance (smallest values) for a given subnetwork (N or F with 10, 30, or 50% validation points) and CPI.

1.8.3 Comparison of EnOI and NRCAN gridded interpolated datasets

The CPI dataset created from EnOI was compared with the CPI values estimated from the NRCAN dataset. Although NRCAN is constructed by interpolation of station records and therefore should adequately reproduce CPI values at sites where stations are located, two major issues must be reminded regarding interpolated gridded datasets. First, station density has an impact on the estimated precipitation suggesting possible large errors in gridded precipitation products for sites far from station network when station density is low (Gervais *et al.*, 2014). Secondly, representativeness error defined by Tustison *et al.* (2001) as “the errors in representing data (i.e., either model output or observations) at a scale other than their own inherent scale” may as well be influenced by the interpolating procedure. Besides, as was pointed out in section 1.3, some stations used in this study were not considered when creating the NRCAN dataset. The NRCAN dataset ($0.083^\circ \times 0.083^\circ$) has better spatial resolution than EnOI ($0.312^\circ \times 0.312^\circ$) and therefore it was necessary to remap NRCAN dataset to the EnOI grid resolution in order to do a grid-to-grid comparison. A conservative remapping method (Jones, 1999) was used.

Figure 1.12 shows maps of EnOI and NRCan mean CPI values over the 1980-2009 period and the corresponding grid-to-grid differences between EnOI and NRCan values for three specific indices (PRmean, PR1mm, and RX1day).

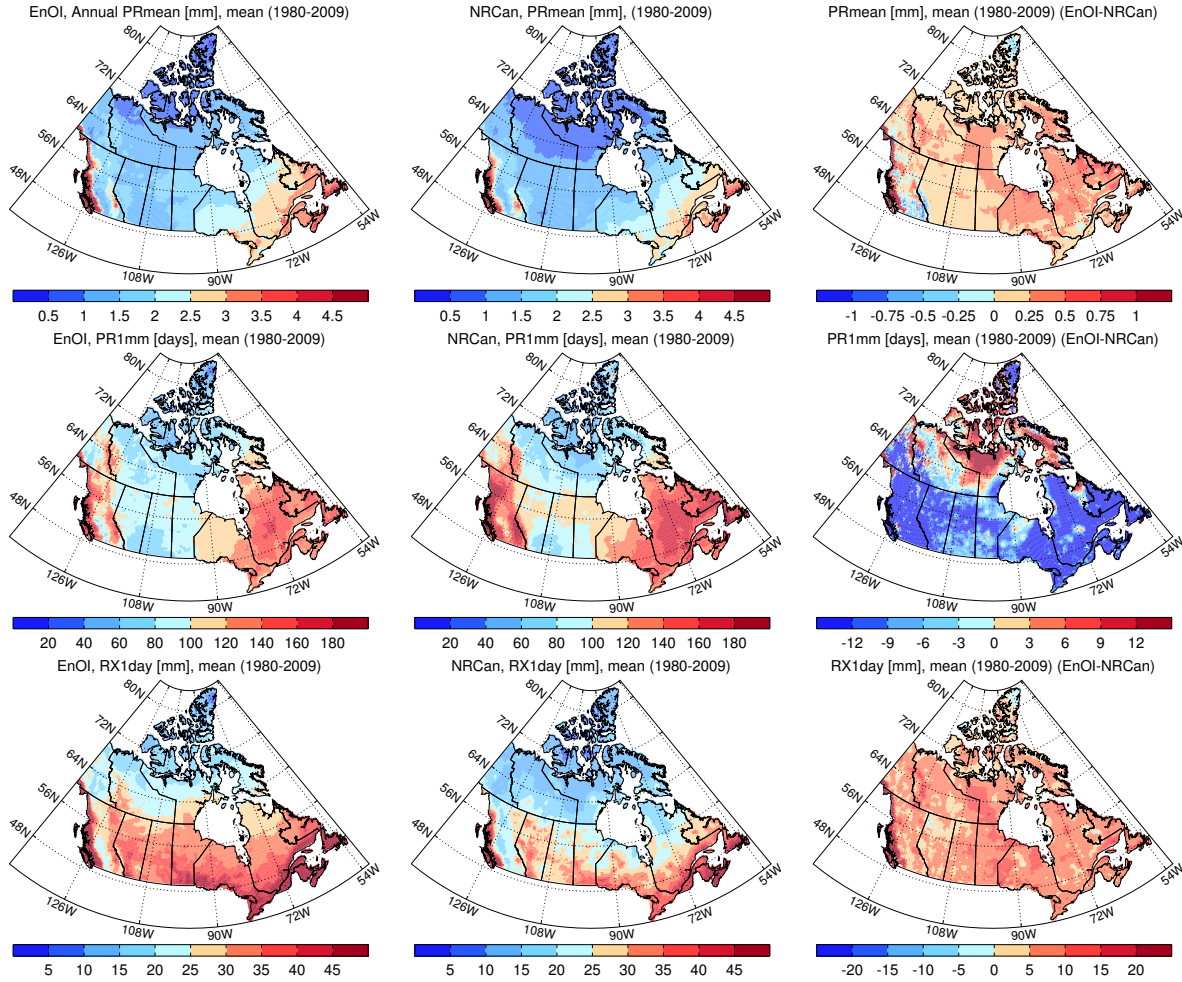


Figure 1.12 – Mean PRmean (first line), PR1mm (second line), and RX1day (third line) values over the 1980-2009 period as estimated from the EnOI (left column), NRCan datasets (middle column), and corresponding grid-to-grid differences between EnOI and NRCan (right column).

The smallest differences between EnOI and NRCan datasets were observed for PRmean (in general smaller than 0.5 mm/day) with mostly smaller values for NRCan dataset. PR1mm values for EnOI are globally lower than NRCan values (especially in the center and south part of the study area). In the northern part, NRCan also displays a small area with lowest values for this index. Lower RX1day values are observed for NRCan for almost the entire domain and this could probably be attributed to the interpolation process used to create this dataset which smooths out RX1day values.

Annual CPI values over the reference period estimated from the EnOI and NRCan datasets were then compared to corresponding station values. Figure 1.13 shows the corresponding scatter plots for these three indices.

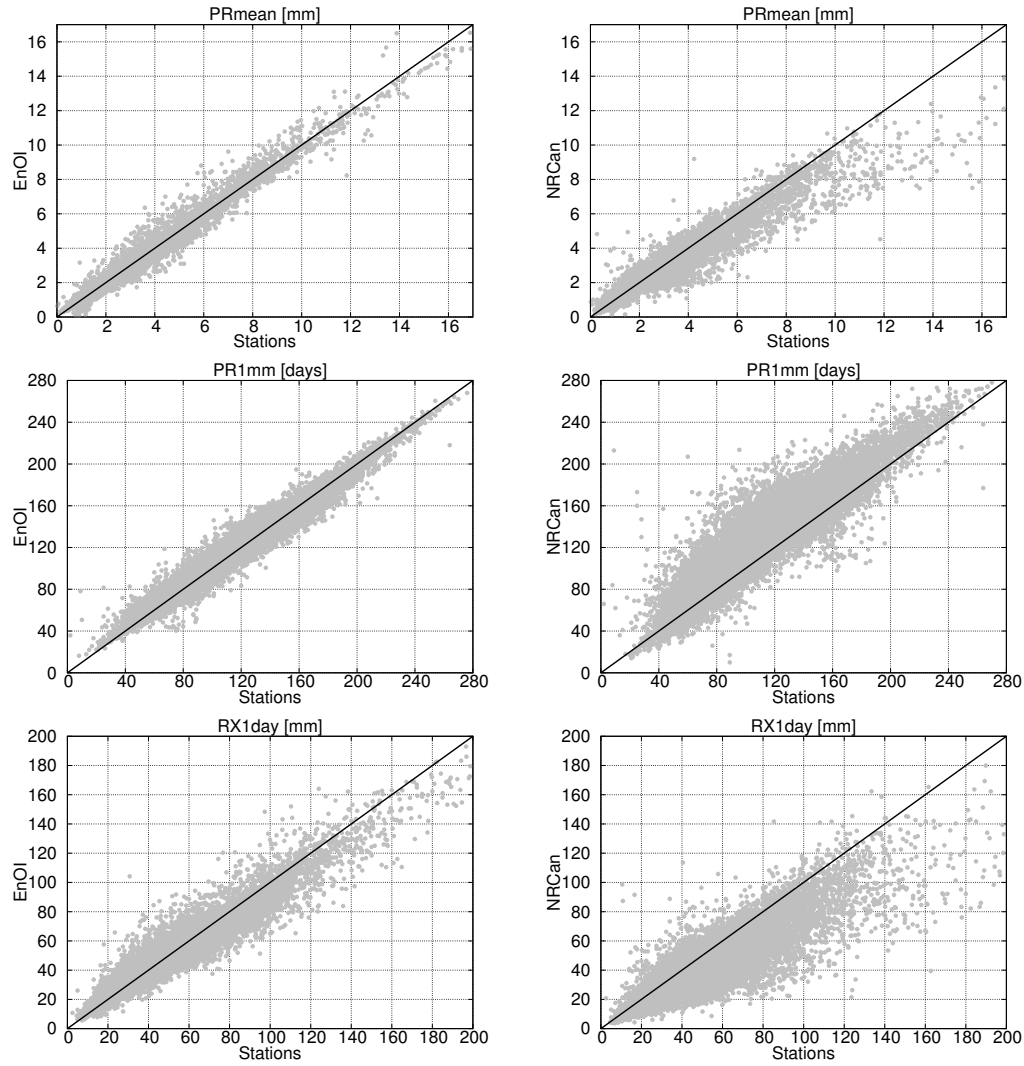


Figure 1.13 – Scatter plots of PRmean (first line), PR1mm (second line), and RX1day (third line) values estimated from EnOI (left column) and NRCan (right column) dataset against the corresponding observed values.

This figure shows that NRCan lightly underestimates observed PRmean values but underestimations are much more important for RX1day especially at sites with large observed RX1day values. Finally PR1mm estimated from NRCan overestimates corresponding station values. Biases are much smaller for EnOI ensemble for all three indices.

Figure 1.14 shows a Taylor diagram with the comparison of EnOI and NRCAN CPI values at observation sites. The diagram clearly shows that EnOI dataset outperforms the NRCAN dataset for each CPI.

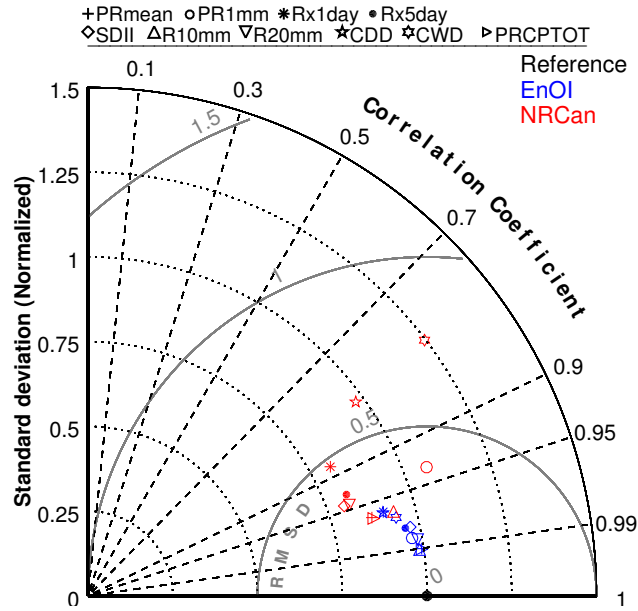


Figure 1.14 – Taylor diagram comparing annual CPI values at station sites estimated with NRCAN gridded dataset (red symbols) and after applying EnOI (blue symbols).

These results show that the CPI estimated from the EnOI dataset, obtained through the combination of the four reanalyses with the observational dataset, outperformed NRCAN. Also, these results show that NRCAN has to be used with caution for extreme indices in Canada.

1.9 Summary and conclusions

It is important to have an adequate reference climate when assessing the impact of climate change and comparing various adaptation strategies. However it can be challenging in remote regions historically poorly covered by station networks. Reanalyses represent an interesting option for such regions. Ground precipitation records being not assimilated, precipitation simulated by reanalysis may however have large biases. Combining reanalysis and station records could therefore provide an improvement in the spatial representation of precipitation. In this study, the off-line data assimilation techniques (OI and EnOI) were applied to define a historical climate for Canada through the combination of observational and reanalysis datasets. The OI method was applied by combining each reanalysis independently with the available station records while the EnOI method combines station records with the information provided by all reanalyses. A period of 30 years (1980-2009) was considered. Annual series for ten Climate Precipitation Indices (CPI) were estimated for each OI and EnOI datasets.

In order to verify the performance of the dataset created after applying each assimilation method, a statistical analysis was first performed comparing the estimated annual precipitation indicator values over the 1980-2009 period at each observational sites to the corresponding nearest grid-point values estimated through OI or EnOI. Results showed that reanalysis performance strongly depends on considered indicators. For instance, indicators displaying the best overall performance were annual mean precipitation (PRmean) and the annual total precipitation from days with precipitation ≥ 1 mm (PRCPTOT), while the indicator showing the poorest performance was the annual maximum number of consecutive days with precipitation ≥ 1 mm (CWD). JRA55 reanalysis globally outperformed the other reanalyses while CFSR and MERRA were the ones with the poorest global performance. As expected, OI and EnOI led to large reductions in the differences between observed and estimated values at station sites.

Cross-validation was then applied in order to evaluate the performance at validation sites not included into the calibration process. Two different subnetworks, *Near* and *Far*, were defined based on the average distance between calibration and validation sites. Validation networks including 10, 30, and 50% of the stations were considered. A total of 60 subnetworks per dataset (six for each of the 10 CPI, three *Near* and three *Far*) were considered, showing improvements in the estimated OI CPI annual series compared to their respective reanalysis. These improvements were larger for CPI

estimated from CFSR and MERRA than the other reanalyses after applying OI. Among the data obtained by OI, CPI estimated from JRA55 and ERAI presented the best performance while those estimated from MERRA and CFSR showed the lowest. Finally, EnOI dataset also outperformed OI applied to each reanalysis.

CPI annual series estimated from the EnOI dataset were finally compared with the corresponding values estimated from the NRCan dataset. A grid-to-grid comparison was carried out for three specific indices: annual mean daily precipitation (PRmean); number of wet days (PR1mm); annual maximum 1-day precipitation (RX1day). Differences between each of these datasets and observed values remained small for PRmean while NRCan globally overestimated PR1mm values and underestimated RX1day values. Smoothing effect due to the interpolation method may partly explained these biases. These results show that NRCan should be used with caution for extreme indicators such as RX1day.

The proposed dataset provides a unique reference climate dataset for ten precipitation indices covering all Canada for the 1980-2009 period. It combines the main reanalyses actually available that were bias corrected with available station records. Such gridded dataset increases our ability to characterize the past climate over regions poorly covered with station networks and it is an important contribution to the climate change impact studies as well as the development of adaptation strategies. Future work will look at the extension of the presented period in this study to include the recent years, as well as incorporate other datasets in the assimilation process (e.g., satellite data). Different comparison analysis will be carry out with other available datasets for the recent years (e.g., the Canadian Precipitation Analysis - CaPa). Besides, application of these methods to daily precipitation series is planned, enabling a possible application to hydrology.

References

- Alexander LV, Zhang X, Peterson TC, Caesar J, Gleason B, Klein Tank AMG, Haylock M, Collins D, Trewin B, Rahimzadeh F, Tagipour A, Rupa Kumar K, Revadekar J, Griffiths G, Vincent L, Stephenson DB, Burn J, Aguilar E, Brunet M, Taylor M, New M, Zhai P, Rusticucci M & Vazquez-Aguirre JL (2006). Global observed changes in daily climate extremes of temperature and precipitation. *Journal of Geophysical Research: Atmospheres*, 111:D05109. DOI:10.1029/2005JD006290.
- Alexeev VA, Esau I, Polyakov IV, Byam SJ & Sorokina S (2011). Vertical structure of recent arctic warming from observed data and reanalysis products. *Climatic Change*, 111(2):215–239. DOI:10.1007/s10584-011-0192-8.
- Bajamngigni Gbambie AS, Poulin A, Boucher MA & Arsenault R (2017). Added Value of Alternative Information in Interpolated Precipitation Datasets for Hydrology. 18(1):247–264. DOI:10.1175/JHM-D-16-0032.1.
- Bertino L, Evensen G & Wackernagel H (2007). Sequential Data Assimilation Techniques in Oceanography. *International Statistical Review*, 71(2):223–241. DOI:10.1111/j.1751-5823.2003.tb00194.x.
- Bosilovich MG, Chen J, Robertson FR & Adler RF (2008). Evaluation of Global Precipitation in Reanalyses. *Journal of Applied Meteorology and Climatology*, 47(9):2279–2299. DOI:10.1175/2008JAMC1921.1.
- Buell CE (1972). Correlation Functions for Wind and Geopotential on Isobaric Surfaces. *Journal of Applied Meteorology*, 11:51–59. DOI:10.1175/1520-0450(1972)011<0051:CFFWAG>2.0.CO;2.
- Candiani G, Carnevale C, Finzi G, Pisoni E & Volta M (2013). A comparison of reanalysis techniques: Applying optimal interpolation and ensemble kalman filtering to improve air quality monitoring at mesoscale. *Science of the Total Environment*, 458-460:7–14. DOI:10.1016/j.scitotenv.2013.03.089.
- Cao L, Hou Y & Qi P (2015). Altimeter significant wave height data assimilation in the South China Sea using Ensemble Optimal Interpolation. *Chinese Journal of Oceanology and Limnology*, 33(5):1309–1319. DOI:10.1007/s00343-014-4252-6.
- Cherubini T, Ghelli A & Lalaurette F (2002). Verification of Precipitation Forecasts over the Alpine Region Using a High-Density Observing Network. *Weather and Forecasting*, 17(2):238–249. DOI:10.1175/1520-0434(2002)017<0238:VOPFOT>2.0.CO;2.
- Coles S (2001). *An Introduction to Statistical Modeling of Extreme Values*, Springer, London. 208 pages.

- Daley R (1991). *Atmospheric Data Analysis*. Cambridge University Press.
- Dee DP, Uppala SM, Simmons AJ, Berrisford P, Poli P, Kobayashi S, Andrae U, Balmaseda MA, Balsamo G, Bauer P, Bechtold P, Beljaars ACM, van de Berg L, Bidlot J, Bormann N, Delsol C, Dragani R, Fuentes M, Geer AJ, Haimberger L, Healy SB, Hersbach H, Hólm EV, Isaksen L, Kållberg P, Köhler M, Matricardi M, McNally AP, Monge-Sanz BM, Morcrette JJ, Park BK, Peubey C, de Rosnay P, Tavolato C, Thépaut JN & Vitart F (2011). The ERA-Interim reanalysis: configuration and performance of the data assimilation system. *Quarterly Journal of the Royal Meteorological Society*, 137(656):553–597. DOI:10.1002/qj.828.
- DeGaetano AT (2009). Time-Dependent Changes in Extreme-Precipitation Return-Period Amounts in the Continental United States. *Journal of Applied Meteorology and Climatology*, 48(10): 2086–2099. DOI:10.1175/2009JAMC2179.1.
- Diaconescu EP, Mailhot A, Brown R & Chaumont D (2017). Evaluation of CORDEX-Arctic daily precipitation and temperature-based climate indices over Canadian Arctic land areas. *Climate Dynamics*, pages 1–25. DOI:10.1007/s00382-017-3736-4.
- Donat MG, Alexander LV, Yang H, Durre I, Vose R, Dunn RJH, Willett KM, Aguilar E, Brunet M, Caesar J, Hewitson B, Jack C, Klein Tank AMG, Kruger AC, Marengo J, Peterson TC, Renom M, Oria Rojas C, Rusticucci M, Salinger J, Elayah AS, Sekele SS, Srivastava AK, Trewin B, Villarroel C, Vincent LA, Zhai P, Zhang X & Kitching S (2013). Updated analyses of temperature and precipitation extreme indices since the beginning of the twentieth century: The HadEX2 dataset. *Journal of Geophysical Research: Atmospheres*, 118(5):2098–2118. DOI:10.1002/jgrd.50150.
- Donat MG, Sillmann J, Wild S, Alexander LV, Lippmann T & Zwiers FW (2014). Consistency of Temperature and Precipitation Extremes across Various Global Gridded In Situ and Reanalysis Datasets. *Journal of Climate*, 27(13):5019–5035. DOI:10.1175/JCLI-D-13-00405.1.
- Evensen G (2003). The Ensemble Kalman Filter: theoretical formulation and practical implementation. *Ocean Dynamics*, 53(4):343–367. DOI:10.1007/s10236-003-0036-9.
- Fortin V, Roy G, Donaldson N & Mahidjiba A (2015). Assimilation of radar quantitative precipitation estimations in the Canadian Precipitation Analysis (CaPA). *Journal of Hydrology*, 531:296–307. DOI:10.1016/j.jhydrol.2015.08.003.
- Frich P, Alexander LV, DellaMarta P, Gleason B, Haylock M, Tank AMGK & Peterson T (2002). Observed coherent changes in climatic extremes during the second half of the twentieth century. *Climate Research*, 19(3):193–212. DOI:10.3354/cr019193.
- Fu W, She J & Zhuang S (2011). Application of an Ensemble Optimal Interpolation in a North/Baltic Sea model: Assimilating temperature and salinity profiles. *Ocean Modelling*, 40(3-4):227–245. DOI:10.1016/j.ocemod.2011.09.004.
- Gervais M, Tremblay LB, Gyakum JR & Atallah E (2014). Representing Extremes in a Daily Gridded Precipitation Analysis over the United States: Impacts of Station Density, Resolution, and Gridding Methods. *Journal of Climate*, 27(14):5201–5218. DOI:10.1175/JCLI-D-13-00319.1.
- Hartmann D, A.M.G. Klein Tank, M. Rusticucci, L.V. Alexander, S. Brönnimann, Y. Charabi, F.J. Dentener, E.J. Dlugokencky, D.R. Easterling, A. Kaplan, B.J. Soden, P.W. Thorne, M. Wild, P.M. Zhai & 2013: (2013). Observations: Atmosphere and Surface. In: Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report

- of the Intergovernmental Panel on Climate Change [Stocker, T.F., D. Qin, G.-K. Plattner, M. Tignor, S.K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex and P.M. Midgley (eds.)]. *Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA*.
- Hatzaki M, Flocas H, Oikonomou C & Giannakopoulos C (2010). Future changes in the relationship of precipitation intensity in Eastern Mediterranean with large scale circulation. *Advances in Geosciences*, 23:31–36. DOI:10.5194/adgeo-23-31-2010.
- Heim RR (2015). An overview of weather and climate extremes – Products and trends. *Weather and Climate Extremes*, 10:1–9. DOI:10.1016/j.wace.2015.11.001.
- Hintze JL & Nelson RD (1998). Violin Plots: A Box Plot-Density Trace Synergism. *The American Statistician*, 52(2):181–184. DOI:10.1080/00031305.1998.10480559.
- Hutchinson MF, McKenney DW, Lawrence K, Pedlar JH, Hopkinson RF, Milewska E & Papadopol P (2009). Development and Testing of Canada-Wide Interpolated Spatial Models of Daily Minimum–Maximum Temperature and Precipitation for 1961–2003. 48(4):725–741. DOI:10.1175/2008JAMC1979.1.
- Häggmark L, Ivarsson KI, Gollvik S & Olofsson PO (2000). Mesan, an operational mesoscale analysis system. *Tellus A: Dynamic Meteorology and Oceanography*, 52(1):2–20. DOI:10.3402/tellusa.v52i1.12250.
- Ji Q, Zhu X, Wang H, Liu G, Gao S, Ji X & Xu Q (2015). Assimilating operational SST and sea ice analysis data into an operational circulation model for the coastal seas of China. *Acta Oceanologica Sinica*, 34(7):54–64. DOI:10.1007/s13131-015-0691-y.
- Jones PW (1999). First- and Second-Order Conservative Remapping Schemes for Grids in Spherical Coordinates. *Monthly Weather Review*, 127(9):2204–2210. DOI:10.1175/1520-0493(1999)127<2204:FASOCR>2.0.CO;2.
- Kalnay E (2002). *Atmospheric Modeling, Data Assimilation and Predictability*. Cambridge University Press, 1 edition edition.
- Kalnay E, Kanamitsu M, Kistler R, Collins W, Deaven D, Gandin L, Iredell M, Saha S, White G, Woollen J, Zhu Y, Leetmaa A, Reynolds R, Chelliah M, Ebisuzaki W, Higgins W, Janowiak J, Mo KC, Ropelewski C, Wang J, Jenne R & Joseph D (1996). The NCEP/NCAR 40-Year Reanalysis Project. *Bulletin of the American Meteorological Society*, 77(3):437–471. DOI:10.1175/1520-0477(1996)077<0437:TNYRP>2.0.CO;2.
- Kharin VV, Zwiers FW, Zhang X & Hegerl GC (2007). Changes in Temperature and Precipitation Extremes in the IPCC Ensemble of Global Coupled Model Simulations. *Journal of Climate*, 20(8):1419–1444. DOI:10.1175/JCLI4066.1.
- Kimoto M, Yoshikawa I & Ishii M (1997). An ocean data assimilation system for climate monitoring. *Journal of the Meteorological Society of Japan*, 75(1 B):471–487.
- Kobayashi S, Ota Y, Harada Y, Ebita A, Moriya M, Onoda H, Onogi K, Kamahori H, Kobayashi C, Endo H, Miyaoka K & Takahashi K (2015). The JRA-55 Reanalysis: General Specifications and Basic Characteristics. *Journal of the Meteorological Society of Japan. Ser. II*, 93(1):5–48. DOI:10.2151/jmsj.2015-001.

- Lachance-Cloutier S, Turcotte R & Cyr JF (2017). Combining streamflow observations and hydrologic simulations for the retrospective estimation of daily streamflow for ungauged rivers in southern Quebec (Canada). *Journal of Hydrology*, 550:294–306. DOI:10.1016/j.jhydrol.2017.05.011.
- Lespinas F, Fortin V, Roy G, Rasmussen P & Stadnyk T (2015). Performance Evaluation of the Canadian Precipitation Analysis (CaPA). *Journal of Hydrometeorology*, 16(5):2045–2064. DOI:10.1175/JHM-D-14-0191.1.
- Lindsay R, Wensnahan M, Schweiger A & Zhang J (2014). Evaluation of Seven Different Atmospheric Reanalysis Products in the Arctic. *Journal of Climate*, 27(7):2588–2606. DOI:10.1175/JCLI-D-13-00014.1.
- Lorenz C & Kunstmann H (2012). The Hydrological Cycle in Three State-of-the-Art Reanalyses: Intercomparison and Performance Analysis. *Journal of Hydrometeorology*, 13(5):1397–1420. DOI:10.1175/JHM-D-11-088.1.
- Mahfouf JF, Brasnett B & Gagnon S (2007). A Canadian precipitation analysis (CaPA) project: Description and preliminary results. *Atmosphere-Ocean*, 45(1):1–17. DOI:10.3137/ao.v450101.
- Matsikaris A, Widmann M & Jungclaus J (2015). On-line and off-line data assimilation in palaeoclimatology: a case study. *Clim. Past*, 11(1):81–93. DOI:10.5194/cp-11-81-2015.
- McKenney DW, Hutchinson M, Papadopol P, Lawrence K, Pedlar J, Campbell K, Milewska E, Hopkinson R, Price D & Owen T (2011). Customized spatial climate models for North America. *Bulletin of American Meteorological Society-BAMS December*, 92(12):1611–1622. DOI:10.1175/2011BAMS3132.1.
- Mekis E & Vincent L (2011). An Overview of the Second Generation Adjusted Daily Precipitation Dataset for Trend Analysis in Canada. *Atmosphere-Ocean*, 49(2):163–177. DOI:10.1080/07055900.2011.583910.
- Mignac D, Tanajura C, Santana A, Lima L & Xie J (2015). Argo data assimilation into HYCOM with an EnOI method in the Atlantic Ocean. *Ocean Science*, 11(1):195–213. DOI:10.5194/os-11-195-2015.
- Mishra AK & Singh VP (2010). Changes in extreme precipitation in Texas. *Journal of Geophysical Research: Atmospheres*, 115:D14106. DOI:10.1029/2009JD013398.
- Nie S, Wu T, Luo Y, Deng X, Shi X, Wang Z, Liu X & Huang J (2016). A strategy for merging objective estimates of global daily precipitation from gauge observations, satellite estimates, and numerical predictions. *Advances in Atmospheric Sciences*, 33(7):889–904. DOI:10.1007/s00376-016-5223-y.
- Oke P, Brassington G, Griffin D & Schiller A (2010). Ocean data assimilation: A case for ensemble optimal interpolation. *Australian Meteorological and Oceanographic Journal*, 59(SPECIAL ISSUE):67–76.
- Oke PR, Sakov P & Corney SP (2006). Impacts of localisation in the EnKF and EnOI: experiments with a small model. *Ocean Dynamics*, 57(1):32–45. DOI:10.1007/s10236-006-0088-8.

- Panthou G, Mailhot A, Laurence E & Talbot G (2014). Relationship between Surface Temperature and Extreme Rainfalls: A Multi-Time-Scale and Event-Based Analysis. *Journal of Hydrometeorology*, 15(5):1999–2011. DOI:10.1175/JHM-D-14-0020.1.
- Panthou G, Vischel T, Lebel T, Blanchet J, Quantin G & Ali A (2012). Extreme rainfall in West Africa: A regional modeling. *Water Resources Research*, 48(8):W08501. DOI:10.1029/2012WR012052.
- Peterson T, Folland C, Gruza G, Hogg W, Mokssit A & Plummer N (2001). Report on the Activities of the Working Group on Climate Change Detection and Related Rapporteurs 1998-2001. *WMO, Rep. WCDMP-47, WMO-TD 1071, Geneve, Switzerland*, 143 pages.
- Qi P & Cao L (2015). Establishment and tests of EnOI assimilation module for WAVEWATCH III. *Chinese Journal of Oceanology and Limnology*, 33(5):1295–1308. DOI:10.1007/s00343-015-4282-8.
- Rahimpour V, Zeng Y, Mannaerts CM & Su ZB (2016). Attributing seasonal variation of daily extreme precipitation events across The Netherlands. *Weather and Climate Extremes*, 14:56–66. DOI:10.1016/j.wace.2016.11.003.
- Rajczak J, Pall P & Schär C (2013). Projections of extreme precipitation events in regional climate simulations for Europe and the Alpine Region. *Journal of Geophysical Research: Atmospheres*, 118(9):3610–3626. DOI:10.1002/jgrd.50297.
- Ratheesh S, Sharma R & Basu S (2014). An EnOI assimilation of satellite data in an Indian Ocean circulation model. *IEEE Transactions on Geoscience and Remote Sensing*, 52(7):4106–4111. DOI:10.1109/TGRS.2013.2279606.
- Ren L & Hartnett M (2017). Sensitivity Analysis of a Data Assimilation Technique for Hindcasting and Forecasting Hydrodynamics of a Complex Coastal Water Body. *Comput. Geosci.*, 99:81–90. DOI:10.1016/j.cageo.2016.10.012.
- Rienecker MM, Suarez MJ, Gelaro R, Todling R, Bacmeister J, Liu E, Bosilovich MG, Schubert SD, Takacs L, Kim GK, Bloom S, Chen J, Collins D, Conaty A, da Silva A, Gu W, Joiner J, Koster RD, Lucchesi R, Molod A, Owens T, Pawson S, Pegion P, Redder CR, Reichle R, Robertson FR, Ruddick AG, Sienkiewicz M & Woollen J (2011). MERRA: NASA's Modern-Era Retrospective Analysis for Research and Applications. *Journal of Climate*, 24(14):3624–3648. DOI:10.1175/JCLI-D-11-00015.1.
- Rusu L & Guedes Soares C (2016). Comparison of various data assimilation methods to improve the wave predictions in the Portuguese coastal environment. volume 2, pages 1087–1096.
- Saha S, Moorthi S, Pan HL, Wu X, Wang J, Nadiga S, Tripp P, Kistler R, Woollen J, Behringer D, Liu H, Stokes D, Grumbine R, Gayno G, Wang J, Hou YT, Chuang HY, Juang HMH, Sela J, Iredell M, Treadon R, Kleist D, Van Delst P, Keyser D, Derber J, Ek M, Meng J, Wei H, Yang R, Lord S, Van Den Dool H, Kumar A, Wang W, Long C, Chelliah M, Xue Y, Huang B, Schemm JK, Ebisuzaki W, Lin R, Xie P, Chen M, Zhou S, Higgins W, Zou CZ, Liu Q, Chen Y, Han Y, Cucurull L, Reynolds RW, Rutledge G & Goldberg M (2010). The NCEP Climate Forecast System Reanalysis. *Bulletin of the American Meteorological Society*, 91(8):1015–1057. DOI:10.1175/2010BAMS3001.1.
- Santos EB, Lucio PS & Silva CMSe (2016). Estimating return periods for daily precipitation extreme events over the Brazilian Amazon. *Theoretical and Applied Climatology*, 126(3):585–595. DOI:10.1007/s00704-015-1605-9.

Sen Z (2009). *Spatial Modeling Principles in Earth Sciences*. Springer Netherlands.

Seneviratne S, N. Nicholls, D. Easterling, C.M. Goodess, S. Kanae, J. Kossin, Y. Luo, J. Marengo, K. McInnes, M. Rahimi, M. Reichstein, A. Sorteberg, C. Vera & X. Zhang (2012). Changes in climate extremes and their impacts on the natural physical environment. In: Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation [Field, C.B. and V. Barros and T.F. Stocker and D. Qin and D.J. Dokken and K.L. Ebi and M.D. Mastrandrea and K.J. Mach and G.-K. Plattner and S.K. Allen and M. Tignor and P.M. Midgley (eds.)]. A Special Report of Working Groups I and II of the Intergovernmental Panel on Climate Change (IPCC). *Cambridge University Press, Cambridge, UK, and New York, NY, USA*, pp. 109-230.

Sillmann J, Kharin VV, Zhang X, Zwiers FW & Bronaugh D (2013). Climate extremes indices in the CMIP5 multimodel ensemble: Part 1. Model evaluation in the present climate. *Journal of Geophysical Research: Atmospheres*, 118(4):1716–1733. DOI:10.1002/jgrd.50203.

Soci C, Bazile E, Besson F & Landelius T (2016). High-resolution precipitation re-analysis system for climatological purposes. *Tellus A: Dynamic Meteorology and Oceanography*, 68(1):29879. DOI:10.3402/tellusa.v68.29879.

Stocker T, D. Qin, G.-K. Plattner, M. Tignor, S.K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex & P.M. Midgley (eds.) (2013). IPCC, 2013: Summary for Policymakers. In: Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. *Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA*.

Tang Y, Chai T, Pan L, Lee P, Tong D, Kim HC & Chen W (2015). Using optimal interpolation to assimilate surface measurements and satellite AOD for ozone and PM_{2.5}: A case study for July 2011. *Journal of the Air and Waste Management Association*, 65(10):1206–1216. DOI:10.1080/10962247.2015.1062439.

Taylor KE (2001). Summarizing multiple aspects of model performance in a single diagram. *Journal of Geophysical Research: Atmospheres*, 106:7183–7192. DOI:10.1029/2000JD900719.

Tebaldi C, Hayhoe K, Arblaster JM & Meehl GA (2006). Going to the Extremes. *Climatic Change*, 79(3):185–211. DOI:10.1007/s10584-006-9051-4.

Tustison B, Harris D & Foufoula-Georgiou E (2001). Scale issues in verification of precipitation forecasts. *Journal of Geophysical Research: Atmospheres*, 106:11775–11784. DOI:10.1029/2001JD900066.

Vihma T, Pirazzini R, Fer I, Renfrew IA, Sedlar J, Tjernström M, Lüpkes C, Nygård T, Notz D, Weiss J, Marsan D, Cheng B, Birnbaum G, Gerland S, Chechin D & Gascard JC (2014). Advances in understanding and parameterization of small-scale physical processes in the marine Arctic climate system: a review. *Atmos. Chem. Phys.*, 14(17):9403–9450. DOI:10.5194/acp-14-9403-2014.

Wang XL, Xu H, Qian B, Feng Y & Mekis E (2017). Adjusted Daily Rainfall and Snowfall Data for Canada. *Atmosphere-Ocean*, 55(3):155–168. DOI:10.1080/07055900.2017.1342163.

Westra S, Fowler HJ, Evans JP, Alexander LV, Berg P, Johnson F, Kendon EJ, Lenderink G & Roberts NM (2014). Future changes to the intensity and frequency of short-duration extreme rainfall. *Reviews of Geophysics*, 52(3):2014RG000464. DOI:10.1002/2014RG000464.

- Zhang X, Alexander L, Hegerl GC, Jones P, Tank AK, Peterson TC, Trewin B & Zwiers FW (2011). Indices for monitoring changes in extremes based on daily temperature and precipitation data. *Wiley Interdisciplinary Reviews: Climate Change*, 2(6):851–870. DOI:10.1002/wcc.147.
- Zhang X & Zwiers FW (2013). Statistical Indices for the Diagnosing and Detecting Changes in Extremes. *Extremes in a Changing Climate*, Springer, Dordrecht, Water Science and Technology Library, pages 1–14. DOI: 10.1007/978-94-007-4479-0_1.
- Zou H, Zhu J, Zhou L, Li P & Ma S (2014). Validation and application of reanalysis temperature data over the Tibetan Plateau. *Journal of Meteorological Research*, 28(1):139–149. DOI:10.1007/s13351-014-3027-5.
- Zwiers FW, Alexander LV, Hegerl GC, Knutson TR, Kossin JP, Naveau P, Nicholls N, Schär C, Seneviratne SI & Zhang X (2013). Climate Extremes: Challenges in Estimating and Understanding Recent Changes in the Frequency and Intensity of Extreme Climate and Weather Events. *Climate Science for Serving Society*, Springer, Dordrecht, pages 339–389. DOI: 10.1007/978-94-007-6692-1_13.