Reviewer #2
Overall Comment:

The manuscript describes a set of multi-model simulation results which are outcomes of the eartH2Observe project. Although there have been several similar data archives, this is still valuable contribution to our community because such dataset, which provides base information for higher level (derivative) researches, should be updated with updates of observations and models. Also, adding additional spread allows us to have a better estimation of the uncertainty in the state-of-the-art assessments in energy-water cycles and associated processes. It provides a fancy interface to access dataset, and most of data points are alive. However, this work has serious shortcomings to be published:

1) Main questions (P3 #6-9): As the authors also referred to previous researches (e.g., Dirmeyer et al., 2006), question (i) is too general, and has been answered and well-known sense already in our community. Of course, it would be no problem to reconfirm this, but current form, at least, would not be appropriate as a main question. Also, this manuscript does not even provide enough evidence to answer this question. Question (ii) is ambiguous, and it is not answered in the manuscript.

First of all we would like to that the reviewer for the effort taken to review our work.

While we fully agree with the reviewer that the first research question has been investigated previously for other datasets we have asked this question for our dataset specifically and also tried to see for which of the parameters the statement holds. We think this dataset is different enough from previous studies to warrant this. The resolution is different, we have a different mix of model types. The ensemble includes similarly to earlier assessments energy balance based land surface models but also leaky bucket type hydrological models and the forcing we used is different which on it's own can have a large effect on the results.

With respect to the second question we refer to our answer 1 to reviewer 1 where we stated we will remove the question as it is only partially answered. The updated section is copied her also for clarity:

> *"In this paper we present the first version of the dataset, which is based on the current state-of-the-art of the contributing modelling systems and will provide a benchmark to evaluate improvements made to the models and forcing data in the coming years. The main goal of this paper is to provide an overview of the multi-model dataset so that it can be evaluated and used in further research. Secondly, we investigate if the ensemble mean in this dataset is superior to the individual models given the diverse set of models, and if so, for which variables."*

We have also reformulated the introduction to better refer to previous (GLDAS).

2) Validation: The authors should put significant additional efforts to provide information on validations. Since this kind of dataset will be utilized to quantify information critical to society such as water resource availability and to investigate variability and interactions of

associated processes, information related with quality of dataset should be carefully provided. Also, without direct comparison to observation, relative comparison (e.g., SNR here) even would not enough to reconfirm question (i) above. As a minimum effort, I recommend author to include basin-wise validation against in-situ discharge and GRACE TWSA, and to add validation in terms of seasonal variations for each variables. If additional tables which has evaluations in such RMSE, CC, NS are provide, it would be very helpful.

We agree about the importance of validation, but this is not the main objective of the current m/s. The on-line validation reports that are linked from the paper (http://earth2observe.github.io/water-resource-reanalysis-v1/results/ilamb.html) do contain a validation with GRACE for the biomes we used. The strong point of the ilamb system we used is that is provides a consistent presentation of the verification against a diverse set of variables.

- *Furthermore, a pre-existing eartH2Observe validation report based on GRACE and snow depth observations globally and for different regions (not using the ilamb system) will be provided as supplementary material (attached to this answer).*
- *At the basin level a thorough validation of the runoff estimates of the models using discharge observations has already been performed by Beck et al. (2016).*

Therefore, I recommend the editorial board to ask the authors “major revision” to publish this manuscript in Earth Syst. Sci. Data.

Specific Comments

3) P2 #20-21 : “optimizing” is not a right word. Does it mean bias-correction? Please make sure the WATCH forcing dataset is made by randomly resampling EAR40 and correcting bias referring to observations.

The history of the WATCH forcing data is as follows:

a) WATCH Forcing Data for 1958-2001 was created by bias-correcting ERA40 (so e.g. monthly mean Tair_WFD matches Tair_CRU, AND sub-monthly variability matches the ERA40 reanalysis variability and hence reality).
b) WATCH Forcing Data for 1901-1959 was created by using random years of ERA40 and bias correcting as before (so e.g. monthly mean Tair_WFD matches Tair_CRU, BUT sub-monthly variability is realistic but does NOT match the actual weather).
c) During EMBRACE WFDEI for 1979-2014 was created using WFD (1958-2001) methodology but based on ERA-Interim instead of ERA40.

To make this more clear (and remove “optimising”) we have reformulated the section as follows:

“The WATCH programme (Harding et al., 2011) used the WATCH Forcing Data which, for Jan 1958 to Dec 2001, was created by bias correcting ERA40 reanalysis data

(Uppala et al., 2005) using gridded in situ meteorological observations (Weedon et al., 2011). For Jan 1901 to Dec 1957 the WATCH Forcing Data applied the same system of bias correction, but applied to randomly selected years of the ERA40 1958-2001 data (Weedon et al., 2011). During the EMBRACE programme the WATCH Forcing Data (1958-2001) methodology was applied to the more recent ERA-Interim reanalysis (Dee et al., 2011) to create the WFDEI (Weedon et al 2014)."

4) P3 #7-9 : Questions should be revised.

We have reformulated question 1 and the main goal of the paper (present the dataset). The second question has been removed as it is indeed only partially answered. See answer to question 1 for the updated text.

5) P4 Table 1 : Is Bulck should be Bulk? It would be nicer to include reference information for each model

Indeed, this should be Bulk. The main reference information for each model is included in the bullet list with all the models starting on page 3.

6) P6 #19-21 : Please put additional table to show models to availabilities of variables which should be a necessary information to interpret performance of ensemble mean and spread.

This information can be derived from the on-line tables and the subset of those tables in Appendix 2. Together with the new table 4 (see below) we hope to have provided enough information

7) P7 Table 2 : Please add meta-info (e.g., standard name) in cf-convention.

Standard names have been added to the table

8) P7 #9-11 : The mirror of THREDDS is not accessible.

The link http://al-tc002.xtr.deltares.nl:8080/thredds/catalog.html (accessed Fed 6 2017) was unfortunately blocked by a firewall. This has now been corrected and it works as expected and points to the root of the THREDDS server.

9) P8 Table 3 : This is incomplete. At least, it needs to add variables used in validation such as SWE and SC. Also, it is necessary to know each models' averaging depth.

As stated in the caption of the table the list is indeed a selection and thus incomplete. We have chosen a subset of variables as the list of available output differs greatly between the models. SWE is present in the table but SC (SnowFrac) and CanopyInt have been omitted and will be added. We will also add a separate table that includes the averaging depth for the soil moisture used in the validation for each model in the section that describes the soil moisture validation, see below.

Table 6. Averaging depth of Surface moisture and root zone moisture in the models. * using the sixth layer in SoilMoist, ** variable depth supplied with model, *** bucket model, depth=0.970m/(theta_fc-theta_wp), where 0.970m is the capacity of the bucket for SWBM and using the FC parameter for HBV-SIMREG.

	SurfMoist	RootMoist
HTESSEL-CaMa	0.07m	1m
JULES	0.1m	1m
LISFLOOD	-	var**
ORCHIDEE	0.092m*	2m
PCR-GLOBWB	var **	1m
SURFEX-TRIP	0.04m	var**
SWBM	-	var***
W3RA	0.05m	1m
WaterGAP3	-	var **
HBV-SIMREG	var***	var***

10) P10 #24-25 : Why precipitation show different results from the models? I assume this is input dataset and should be identical to models.

All the models were forced by the same precipitation. This statement refers to the comparison of the different precipitation datasets highlighting the large uncertainty of the precipitation variability in these different sources, which was not considered in the earth2observe dataset that only accounts for the multi-model uncertainty. This is further highlighted in the end of the paragraph: "while over the tropical areas some of the multi-model agreement might be underestimating the actual uncertainty by neglecting the driving data uncertainty in the ensemble generation.".
To avoid confusion, the sentence was modified from:
" Comparing these results with the precipitation datasets agreement (Figure 2d and 3d),the large uncertainty in the tropical areas is not reflected in the runoff or root zone soil moisture" to :
"Comparing these results with the precipitation datasets agreement (Figure 2d and 3d), which were not included in the driving data, the large uncertainty in the tropical areas is not reflected in the runoff or root zone soil moisture"

11) P10 #34 : Please provide more information on calculating TWSA and available storage components for each model.

A new table (Table 4) has been added to show the components used to estimate TSW for each model. It is similar to Table 1.1. In the supplementary information. An unformatted version of the table is shown below:

Table 4: Components used in Total Water Storage estimation for each model. The definition of the variables can be found in Table 2

	SWE	*CanopInt*	*SurfStor*	*TotMoist*	*GroundMoist*
HTESSEL-CaMa	*x*	*x*	*x*	*x*	*-*
JULES	*x*	*x*	*-*	*x*	*-*
LISFLOOD	*x*	*-*	*-*	*x*	*x*

ORCHIDEE	*x*	*-*	*x*	*x*	*-*
PCR-GLOBWB	*x*	*x*	*x*	*x*	*x*
SURFEX-TRIP	*x*	*x*	*x*	*x*	*-*
SWBM	*x*	*-*	*-*	*x*	*-*
W3RA	*x*	*-*	*-*	*x*	*x*
WaterGAP3	*x*	*x*	*x*	*-*	*-*
HBV-SIMREG	*x*	*-*	*-*	*x*	*x*

The anomaly was calculated inside the iLamb system by subtracting the mean.

12) P13 #1-2 : Not clear.

We have reformulated this as follows:

> *"Although there are a number of uncertainties associated with TWSA as estimated by GRACE measurements (Long et al., 2014; Riegger et al., 2012) resulting from the uncertainty of the GRACE data itself and the leakage corrections, the results provide an independent mean of evaluating our model results."*

13) P13 #9-10 : It may imply systematic error in the models and/or missing components in analysis. For examples, appropriate treatment of river and groundwater would introduce additional delay and amplitude (e.g., Kim et al. 2009)

Yes, we agree with that and have added a remark at that point and included the Kim reference.

14) P18 Figure 8 : Why it only shows AUST and SEAS? Why not global and other regions/basins? This comment should extend to the other variables. Overall, this manuscript is lacking a universal form and strategy of validations.

We partly agree with this comment. By using the ilamb system and placing all results on-line we have made sure we have set a validation metrics that are open, repeatable and consistent. We have chosen to only show AUST and SEAS in the paper for space reasons but we can surely provide all biomes. We are happy to provide all in an appendix. As said before, all results are also available on-line (http://earth2observe.github.io/water-resource-reanalysis-v1/results/ilamb.html) via the DIO and stored indefinitely by zenodo.

15) P19 #6-11 : This part should be extensively revised with additional previous estimations (e.g., Trenberth et al. 2007; Syed et al. 2009; Rodell et al. 2015) Providing an intercomparison table of water balance components would be convenient to readers.

We have provided a summary of the terrestrial water balance of the model results with the sole aim of placing them next to previous estimates and certainly did not aim to provide a review of the global terrestrial water budget. This said omitting Rodell 2015 was sloppy and

this has been corrected. The section has been extended and rewritten along the following lines

> *" Table 8 presents the results of this study together with a selection of previous studies. Although results are not always directly comparable due to differences in land mask and techniques used current results compare reasonable well with previous estimates. Yearly terrestrial runoff (excluding Antarctica and Greenland) from the ten models ranges between 38652 and 55877 km3/yr with an ensemble mean of 46268 km3/yr. Rodell et al. (2015) presented and optimised estimate of global terrestrial 10 runoff of 45900 km3/yr 4400 km3/yr for the the period 2000-2010. Furthermore, the lower estimates compare well with findings from Clark et al. (2015) (44200 2660 km3/yr) while the ensemble mean compares well with the WATCH-based simulations of 49680 km3/yr (Clark et al., 2015) and the results by Haddeland et al. (2011) (42000 to 66000 km3/yr), but are higher than estimates by van Dijk et al. (2014) (20909 km3/yr, based on 430 basins estimated to cover 90% of global Runoff) and Dai et al. (2009) (37288 km3/yr). The relatively high runoff in the estimates that rely mostly on models such as in this study, may in part be caused by the fact that it can include small islands (Syed et al., 2009) which are not represented in the gauge and GRACE based estimates."*

Global Earth Observation
for integrated water resource assessment

Report on uncertainty characterization of the WP5 WRR tier 1 products based on GRACE products and snow depth data

Marie Minvielle, Bertrand Decharme, Jean-Christophe Calvet
CNRM/GAME, Météo-France, CNRS, UMR 3589, 42 avenue Coriolis, 31057 Toulouse Cedex 1, France

Deliverable No: D4.4 – Report on uncertainty characterization of the WP5 WRR tier 1 products
Ref.: WP4 – Task 4.4 – Issue 1.1

Date: December 2015

Deliverable Title	**D4.4 – Report on uncertainty characterization of the WP5 WRR tier 1 products**
Filename	E2O_D44_v1.2
Contributors	Marie Minvielle (Meteo-France)
	Bertrand Decharme (CNRS)
	Jean-Christophe Calvet (Meteo-France)
Reviewer	Eleanor Blyth (CEH)
Date	10/12/2015

Prepared under contract from the European Commission

Grant Agreement No. 603608
Directorate-General for Research & Innovation (DG Research), Collaborative project, FP7-ENV-2013-two-stage

Start of the project: 01/01/2014
Duration: 48 months
Project coordinator: Stichting Deltares, NL

Dissemination level

X	PU	Public
	PP	Restricted to other programme participants (including the Commission Services)
	RE	Restricted to a group specified by the consortium (including the Commission Services)
	CO	Confidential, only for members of the consortium (including the Commission Services)

1 Introduction

The aim of the WP5 was to produce a multi-model ensemble-based global water resources reanalysis (WRR tier 1 reanalysis) in which state-of-the-art land surface models and global hydrological models are forced by the most-accurate global meteorological forcing provided by atmospheric reanalysis. A part of WP4 is the validation of this WP5 water resources reanalysis, and the specific contribution of Météo-France concerns the terrestrial water storage (Task 4.4) and the snow depth (Task 4.1).

This report presents, for these two variables, the validation of the WRR reanalysis (1979-2012) and compares the models' performances. The report contains a description of the variables to be validated, a description of the observed datasets used for the validation, a presentation of the models which have provided the necessary variables and those finally retained, the results of the terrestrial water storage validation, and results of the snow depth validation. For a complete description of the atmospheric forcing used in the project, the different institutes and models, please refer to the deliverable D5.1 of the project.

2 Observed datasets

2.1 Terrestrial water storage

Terrestrial water storage (TWS) consists of snow and ice, surface water, soil moisture and permafrost, groundwater and vegetation water content. The GRACE satellite mission provides time-variable gravity field solutions which allow direct evaluation of the TWS variations. GRACE data can be used to estimate TWS from basin (Crowley et al. 2006; Seo et al. 2006) to continental scale (Schmidt et al. 2006; Tapley et al. 2004). Other studies have also pointed out the possibility of using GRACE for the estimation of groundwater variations (Rodell et al. 2004; Yeh et al. 2006), ice sheet and glacier mass loss and hydrological fluxes. Our objective in this study, is to use GRACE to evaluate the simulated water storage in the different models of the WRR reanalysis. As shown in previous studies, GRACE can indeed be used to evaluate simulated water storage (Vergnes et Decharme 2012, Alkama et al. 2010, Decharme et al. 2010). Simulated TWS are compared to the GRACE products using the same methodology used by Alkama et al. 2010 and Vergnes et Decharme 2012.

GRACE provides monthly TWS variation estimates based on highly accurate maps of the earth's gravity fields over spatial scales of about 300-400km resolution (Wahr et al. 2004; Swenson et al. 2003). The most recent release (RL05) of 3 GRACE gravity model products were used for the analysis, each one generated by 3 different institutions: the Center for Space Research (CSR at the University of Texas), the Jet Propulsion Laboratory (JPL) and the GeoforschungsZentrum (GFZ). Deriving month-to-month gravity field variations from GRACE observations requires a complex methodology, and many parameter choices. As recommended, we used all three data centers' products, and averaged them. For more details concerning GRACE data, please refer to http://grace.jpl.nasa.gov/data/. For this study, GRACE data from April 2002 to December 2014 were available, but we only used the 2002-2012 period, because 2012 marks the end of the WRR reanalysis. During this period, some months are not available, with the result that 122 months are considered.

2.2 Snow Depth

To perform the evaluation of snow depth in the WRR reanalysis, 4 different sources of daily data exist, in different regions of the world. First, 600 stations over Russia from RIHMI-WDC (All-Russia Research Institute of Hydrometeorological Information - World Data Centre, described in Bulygina et al. 2014) with more than 20 years year-round data, the first records starting in 1874. Over USA, 355 stations over 30°N have been retrieved from the National Climatic Data Center (NCDC, ftp ://ftp.ncdc.noaa.gov/pub/data/ghcn/daily/), with more than 40 years and 350 observations per year, the first records starting in 1889. Data from NCDC over Germany, Netherlands Norway and Sweden are also available (ftp ://ftp.ncdc.noaa.gov/pub/data/ghcn/daily/), 448 stations with more than 40 years, 350 observations per year, the first records starting in 1887. We used also 719 quality controlled stations of daily measurements of Canadian snow depth from Ross Brown (Environment Canada, Brown and Braaten 1998) and NCDC : more than 20 years with 300 observations per year on average, first records starting in 1881, last in 2003, and extension beyond 2003 from NCDC. This ensemble of 2120 stations constitute our base of available observations. As explained in Section 4, not all of these stations will not be included in our analysis.

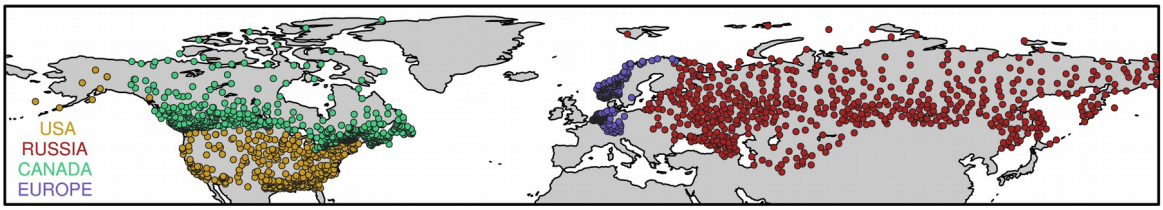

Fig.2.1 Snow depth stations

3 Models

3.1 Terrestrial water storage

In this report, to distinguish the models, the acronyms corresponding to the institutes (first column of Table 1.1) will be used to refer to the model simulations from these institutes.

For comparison to GRACE data, the monthly TWS variations simulated by the models are calculated in terms of anomalies of the sum of total soil moisture (TotMoist), surface water storage (SurfStor, including lakes reservoirs, rivers,..), ground water reservoir (GoundMoist), snow water equivalent (SWE), snow water equivalent intercepted by vegetation (SWEVeg) and total canopy water storage (CanopInt). Ideally, all these variables areneeded to estimate the variations of terrestrial water storage as:

$$\Delta TWS = \Delta TotMoist + \Delta SurfStor + \Delta GroundMoist + \Delta SWE + \Delta SWEVeg + \Delta CanopInt$$

However, a significant disparity exists between the different models taking part in EartH2Observe, and the required variables are not always output (see Table 1).

INSTITUT	MODELE	SWE	CanopInt	SWEVeg	SurfStor	TotMoist	GroundMoist
ECMWF	HTESSEL-CaMa	x	x	-	x	x	-
UNIVU	PCR-GLOBWB	x	x	-	x	x	x
METFR	SURFEX-TRIP	x	x	-	x	x	x
NERC	JULES	x	x	x	-	x	-
JRC	LISFLOOD	x	-	-	-	x	x
CNRS	ORCHIDEE	x	-	-	x	x	-
UNIVK	WaterGAP	x	x	-	x	-	
CSIRO	W3RA	x	-	-	-	x	x
ETH 1	SWBM	x	-	-	-	-	-
2		x	-	-	-	-	-

Table 1.1 Variables required for the calcul of ΔTWS and their availability for each model.

Consequently, depending on the model, the calculated ΔTWS varies and only includes 5, 4 or 3 of these variables. The simulated ΔTWS from the different models, compared below (Section 4), are consequently the sum of different sets of variables. This variation does not create an absolute barrier to performing a global analysis of the WRR reanalysis and having an overview of the ability of the models to reproduce the TWS variations. However, because the calculated TWS for the models are not exactly similar, it is not possible to make a detailed comparison and rank the models. For instance, we decided to not conserve for the study the simulations from ETH, because only SWE was available.

Finally, as mentioned above, the GRACE TWS estimates was first filtered in order to remove noise and errors in the gravity field measurements, which can modify the signal by reducing the seasonal amplitude of the final TWS. To be consistent with the GRACE data, the simulated TWS were smoothed using the same 300 km-width gaussian filter, used by Alkama et al. 2010, which is similar to the filter used for the GRACE products.

3.2 Snow Depth

Snow depth is output by only four of the models (Table 1.2). Here we analyse the daily snow depth at 0.5° resolution for 4 models over the 1979-2012 period. To compare the simulated and observed snow depth, we first have to attribute a station to a model grid point (Brun et al. 2013). For each station, the grid point corresponding to the latitude and longitude of the station is identified. The altitudes of the station and the grid point are compared, and if the difference exceeds 100m, the station is eliminated. Some time-criteria are also introduced, and stations with a lot of missing values can't be used. For example, it is also estimated that if more than 5 consecutive days are missing, the station can't be used to determinate the annual maximum depth snow. Finally, 1424 stations are used to obtain the results given Table 5.1.

INSTITUTE	MODEL	SnowDepth
ECMWF	HTESSEL-CaMa	x
UNIVU	PCR-GLOBWB	-
METFR	SURFEX-TRIP	x
NERC	JULES	x
JRC	LISFLOOD	-
CNRS	ORCHIDEE	x
UNIVK	WaterGAP	-
CSIRO	W3RA	-
ETH 1	SWBM	-
2		-

Table 1.2 Snow depth and its availability for each model.

4 Terrestrial Water Storage – RESULTS

To begin the validation of the observed and simulated ΔTWS, we first present some analysis at the global scale. Figure 4.1 a) shows the monthly global average of simulated and observed ΔTWS, from 2002 to 2012. An important seasonal cycle exists and is represented, for global averaged ΔTWS, Figure 4.1 b). A maximum of the global averaged ΔTWS is observed (black line) during boreal spring and a minimum in autumn. The seasonal cycle is globally well reproduced by the different models. The seasonal cycle of the multi-model mean (grey line) presents a comparable amplitude, but there is a noticeable time lag of one month in the multi-model mean. All models have a tendency to simulate an early seasonal cycle, where the size of the lag varies with the model. It is clear from the timeseries 2002 to 2012 (Fig.4.1) that the seasonal cycle is correctly reproduced, and that the principal differences between GRACE and the models come from the low frequency variability (the trend). In the GRACE product, global averaged ΔTWS tends to decrease. Some models (e.g. METFR) are able to simulate this negative trend, others (e.g. UNIVU) show a contrary positive trend. However, maps of trends have been drawn, and the trends displayed in Fig 4.1 are not representative of a global signal, but are mainly due to some strong trends over few located grid points.

Figure 4.2 shows the spatial distribution of the climatological ΔTWS simulated by the models and estimated by GRACE (first line) from 2002 to 2012, for DJF, MAM, JJA and SON respectively (left to right). For each season, the spatial correlation between the model and GRACE is indicated bottom left. The global spatial pattern of ΔTWS is well represented by the models, particularly in MAM and SON, as already shown in previous studies (Vergnes et al. 2012). Despite good model performance in terms of anomalies, some models significantly underestimate ΔTWS (NERC, JRC, CSIRO or CNRS). But it is important to remember that some models only have 3 of the 6 variables necessary to calculate a perfect ΔTWS. It is concluded here however, that the spatial and seasonal structure are generally acceptable.

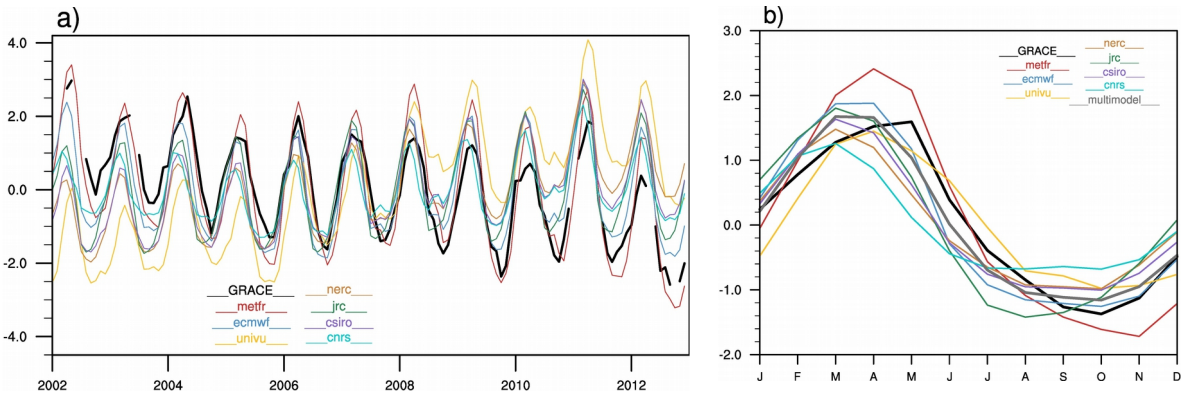

Fig. 4.1 a) Global average of monthly ΔTWS *since 2002, for the GRACE product (black line) and the models (colored lines). b) Annual cycle of* ΔTWS *for the 2002-2012 period. Same color code as a) except the grey line for the multimodel mean.*

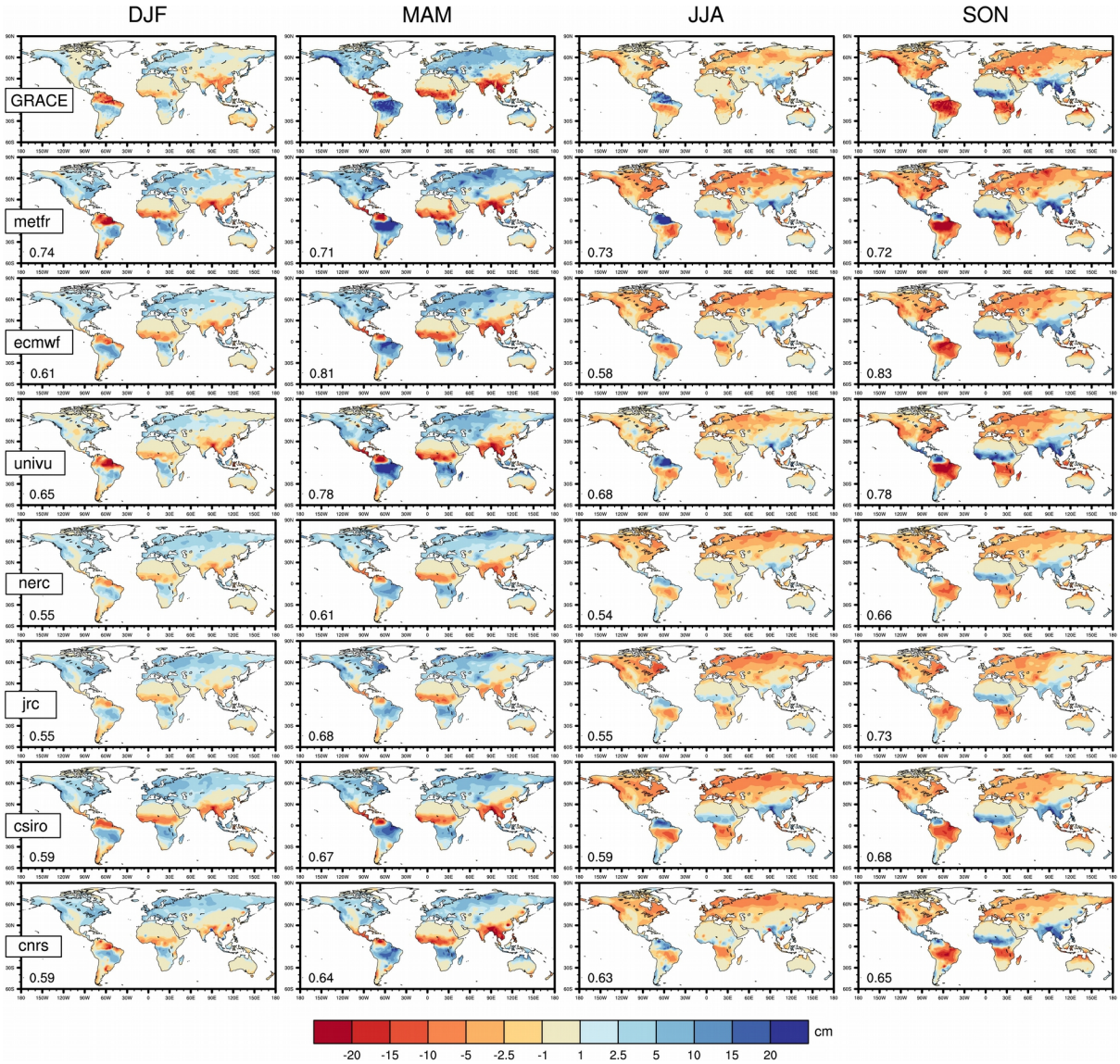

Fig. 4.2 Climatological comparison of the total TWS (cm) between GRACE (top line) and models for (left to right) DJF, MAM,JJA and SON. For each model and each season, the spatial correlation with the TWS GRACE product is indicated bottom left of each map.

The zonal averages of the ΔTWS are given Figure 4.3. In the left column the GRACE product is plotted (black line) with all the models (colours), and in the right column with the multi-model mean with the multi-model spread (in red). The zonal average of the multi-model ΔTWS gives a good estimation of the ΔTWS and is quite similar to the GRACE product, especially in MAM and SON, as noted previously. Although the multi-model mean corresponds closely to the GRACE product, the largest spread between models appears where the ΔTWS is the largest, for instance in the tropical regions.

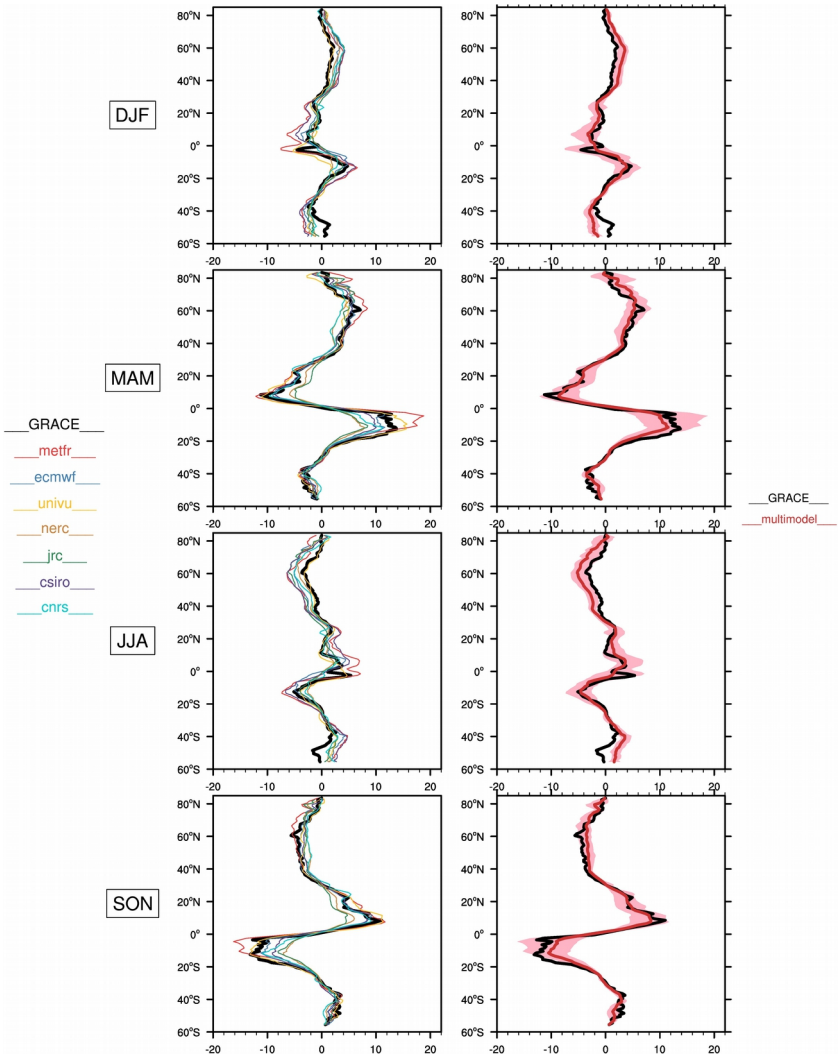

Fig. 4.3 Left column: zonal average of GRACE TWS (black line) and all models (colored lines) for (top to bottom) DJF, MAM, JJA and SON. Right column: GRACE TWS product in black, multimodel average in red, and in pink the multimodel spread.

As seen in the previous figures, the spatial structures of ΔTWS and the values of the anomalies are not uniform and depend of the geographical region. That's why we have chosen to continue the analysis by defining and using 11 geographical boxes. These boxes are represented in Figure 4.4 and correspond to Norther America/Canada (CAN), Western North America (AMO), Eastern North America (AME), Northern Europe (EUN), Southern Europe (EUS), Siberia (SIB), Sahel (SAH), Asia (ASI) , Amazon (AMA), Eastern South America (AMS) and Central Africa (AFR). All results presented thereafter refer to these boxes.

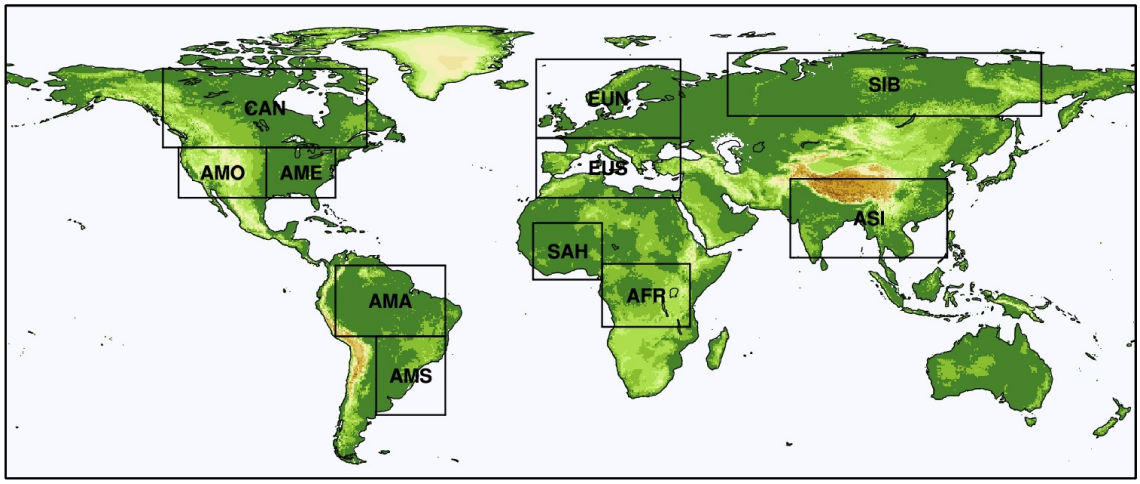

Fig. 4.4 Location of the 11 geographical boxes used for the analysis of TWS.

Figure 4.5 compares the annual cycles of the simulated ΔTWS with the GRACE estimates over the 11 boxes defined Figure 4.4. For some regions (CAN, AMO, AME, AUS, SIB, SAH), as previously discussed in the global mean, the simulated ΔTWS seasonal cycle is shifted and ahead of the GRACE estimate.

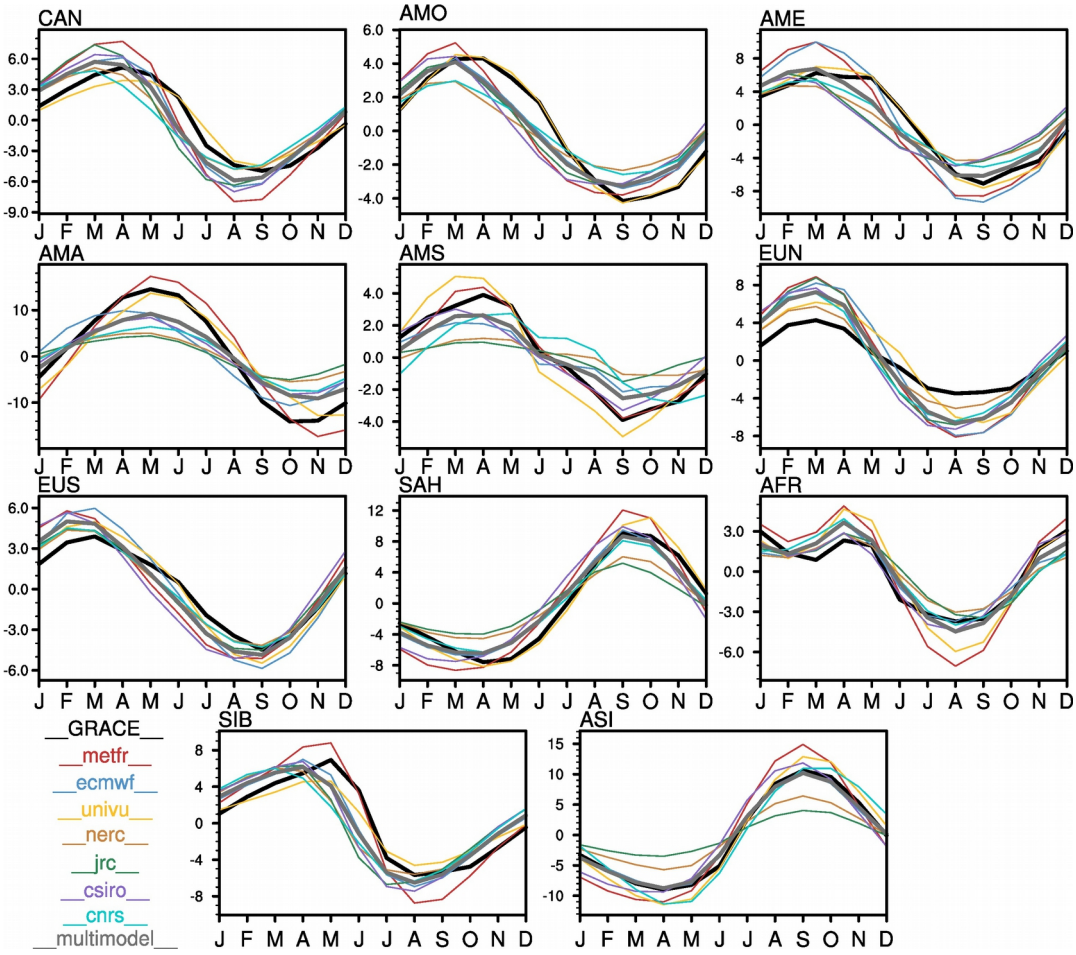

Fig. 4.5 TWS seasonal cycle over the 11 regions previously defined. Black line: GRACE product. Colored lines: simulated TWS.

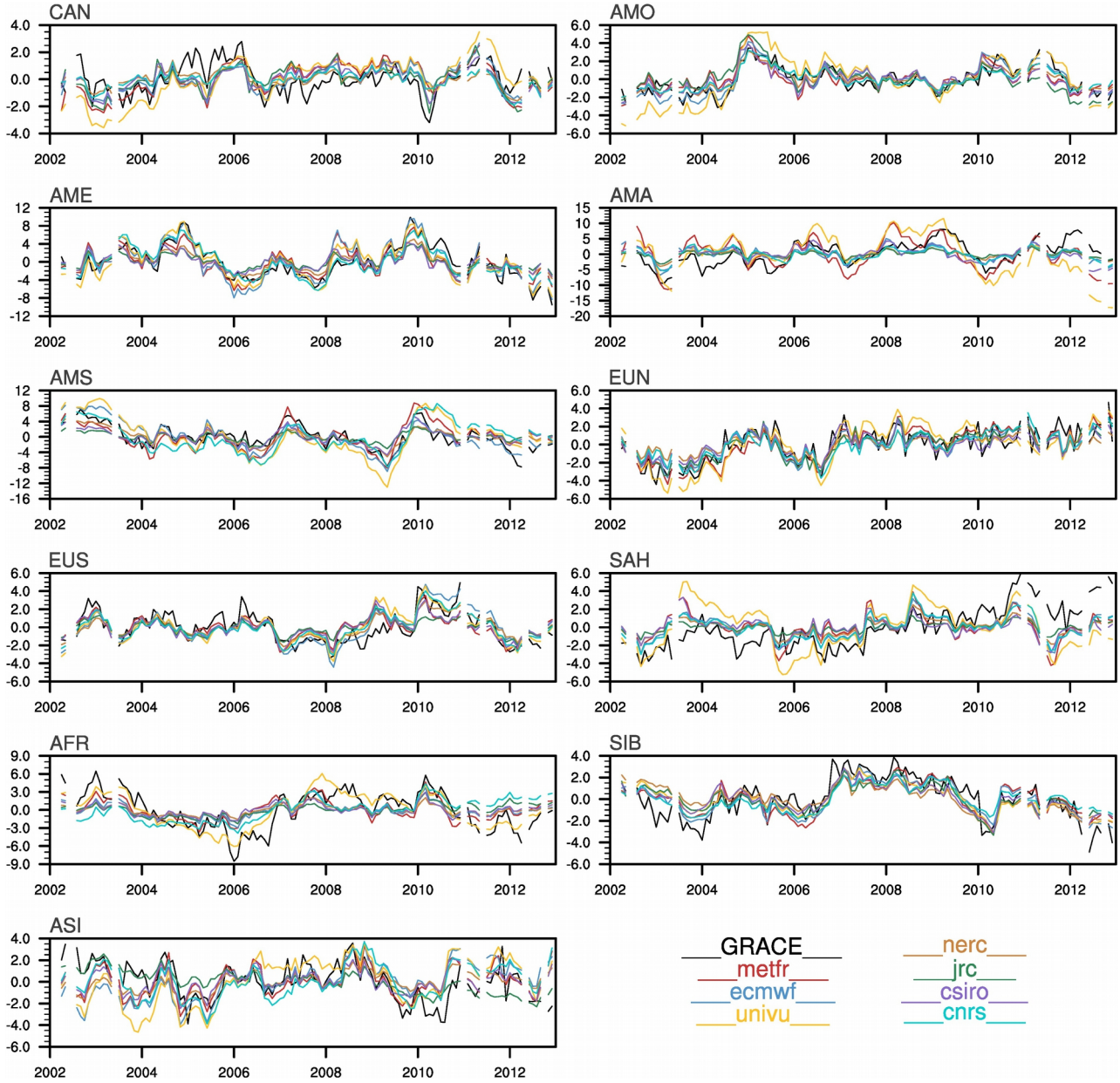

Fig.4.6 Monthly mean ΔTWS anomalies (seasonal cycle removed) over the 11 regions previously defined from 2002 to 2012. Black line: GRACE product. Colored lines: simulated ΔTWS.

A full timeseries of monthly mean ΔTWS anomalies, observed and simulated, averaged over the 11 boxes are represented Figure 4.6. Correlations between the simulated ΔTWS and the GRACE product, corresponding to this Figure 4.6 are grouped in the following table (Table 4.1).
Finally, to complete the analysis of the simulated ΔTWS, we have also calculated the RMSE and the ratio of variance of the monthly anomalies (seasonal cycle removed), between the simulated ΔTWS and the GRACE product, for each model and over each box. Results are grouped together table 4.2 and 4.3.

	CAN	AMO	AME	AMA	AMS	EUN	EUS	SAH	AFR	SIB	ASI
ecmwf	0.43	0.86	0.88	0.60	0.85	0.81	0.81	0.55	0.72	0.84	0.25
univu	0.40	0.87	0.88	0.47	0.72	0.78	0.76	0.34	0.86	0.80	0.36
metfr	0.38	0.70	0.79	0.59	0.73	0.79	0.75	0.30	0.68	0.78	0.49
nerc	0.54	0.87	0.87	0.65	0.82	0.75	0.79	0.51	0.67	0.63	0.35
jrc	0.39	0.63	0.78	0.69	0.69	0.79	0.66	0.59	0.50	0.73	0.62
csiro	0.42	0.74	0.73	0.45	0.73	0.80	0.70	0.14	0.56	0.78	0.48
cnrs	0.48	0.88	0.91	0.63	0.61	0.76	0.79	0.58	0.42	0.77	0.50

Table 4.1 Correlation of the monthly simulated ΔTWS (annual cycle removed) and the GRACE product for each model and each geographical box.

	CAN	AMO	AME	AMA	AMS	EUN	EUS	SAH	AFR	SIB	ASI
ecmwf	1.12	0.82	1.97	3.31	1.95	1.15	1.15	1.97	2.45	1.07	2.14
univu	1.48	1.41	1.88	5.67	3.39	1.43	1.24	2.76	1.61	1.19	2.23
metfr	1.28	1.10	2.35	4.17	2.44	1.23	1.29	2.35	2.39	1.22	1.68
nerc	0.97	0.77	2.31	3.45	2.04	1.40	1.22	2.05	2.73	1.51	1.74
jrc	1.17	1.31	2.79	3.48	2.62	1.19	1.49	2.08	2.82	1.35	1.44
csiro	1.11	1.02	2.79	3.69	2.31	1.23	1.37	2.43	2.77	1.26	1.62
cnrs	1.01	0.74	1.62	3.27	3.21	1.26	1.16	1.91	2.93	1.30	1.72

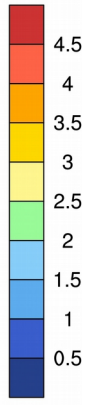

Table 4.2 Same as Table 4.1 for RMSE.

	CAN	AMO	AME	AMA	AMS	EUN	EUS	SAH	AFR	SIB	ASI
ecmwf	0.62	1.15	1.16	0.28	1.21	0.57	0.85	0.27	0.15	0.61	0.80
univu	1.69	2.79	1.08	2.38	2.18	1.35	0.71	1.09	0.81	0.46	1.28
metfr	1.03	0.75	0.70	1.44	1.01	0.87	0.38	0.36	0.29	0.56	0.57
nerc	0.27	0.53	0.26	0.08	0.33	0.22	0.35	0.14	0.06	0.36	0.22
jrc	0.68	1.05	0.16	0.06	0.13	0.58	0.19	0.05	0.09	0.39	0.32
csiro	0.56	0.70	0.23	0.32	0.33	0.39	0.36	0.19	0.09	0.37	0.29
cnrs	0.32	0.53	0.72	0.21	1.36	0.55	0.56	0.30	0.33	0.34	0.72

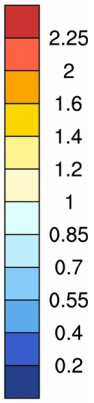

Table 4.3 Same as Table 4.1 for the ratio of variance. A positive (negative) value corresponds to a overestimation (underestimation) of the variance, relative to the GRACE one.

For the correlation and RMSE (Table 4.1 and Table 4.2), results are highly dependent on the considered regions. Models show similar behaviors with relatively high correlations over AMO, AME, AMS, EUN, EUS and SIB, lower over AFR and AMA, and correlations are poor over CAN, SAH and ASI. Obviously, in addition to these general considerations, models are not equals, and some models have consistently better correlation in all regions than others.

Considering the Table with RMSE, results are similar. RMSE is particularly high for some specific regions in any model. This is particularly true for AMA and AME, but also for AMS, SAH, AFR and ASI. For the correlation and RMSE, we can conclude than some models show better results than others, but similarities exist and some regions seem to be more difficult to simulate than others, whatever model is considered.

The last table groups the results for the ratio of variance (Table 4.3). Contrary to correlation and RMSE, the main differences are here due to the model and not the region. The last four models present an significant underestimation of the variance, but it is important to remember that these models are those with the least amount of variables to calculate the ΔTWS.

5 Snow Depth – RESULTS

Various statistical indices were computed in order to evaluate the performance of the different models. More than 1400 stations of daily snow depth have been used to calculate bias, correlation and RMSE for the following variables: annual mean snow depth, number of days per year with snow on the ground, duration of the longest period with continuous snow on the ground, first and last day of the year (since 1st August) with continuous snow. A day with snow on the ground is defined as a day with a snow depth higher than 1cm. The following 4 figures represent maps of bias for these variables, follow-up to the annual cycle of observed and simulated snow depth. Finally, a table groups bias, correlation and RMSE for the four models.

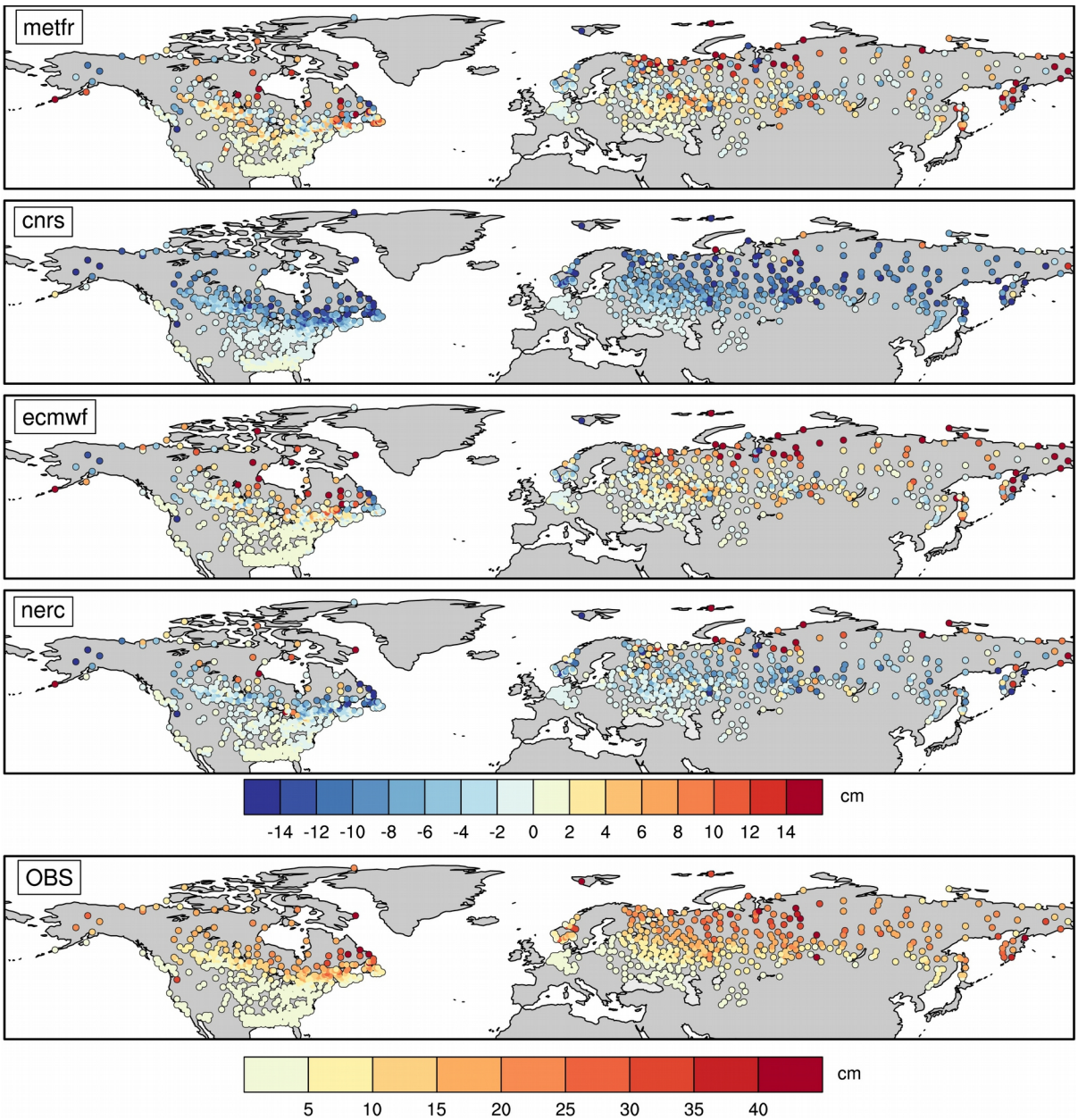

Fig. 5.1 Bias (cm) between the simulated averaged snow depth and the observations over the 1979-2012 period. Bottom : climatological observed values (cm).

Figure 5.1 shows the annual mean snow depth for the stations used to validate the models (bottom), and the maps of bias for the four available models. In terms of annual average, two models (METFR and ECMWF) tend to overestimate the snow depth while others (CNRS and NERC) underestimate it.

Figure 5.2 shows number of days with continuous snow on the ground in the observations (bottom map), defined as days with at least 1cm of snow on the ground, and the bias in the models (4 first maps). METFR, NERC overestimate this duration. NERC strongly overestimate it, especially at highest latitudes. In contrast, CNRS shows a number of days with continuous snow significantly lower than the observations.

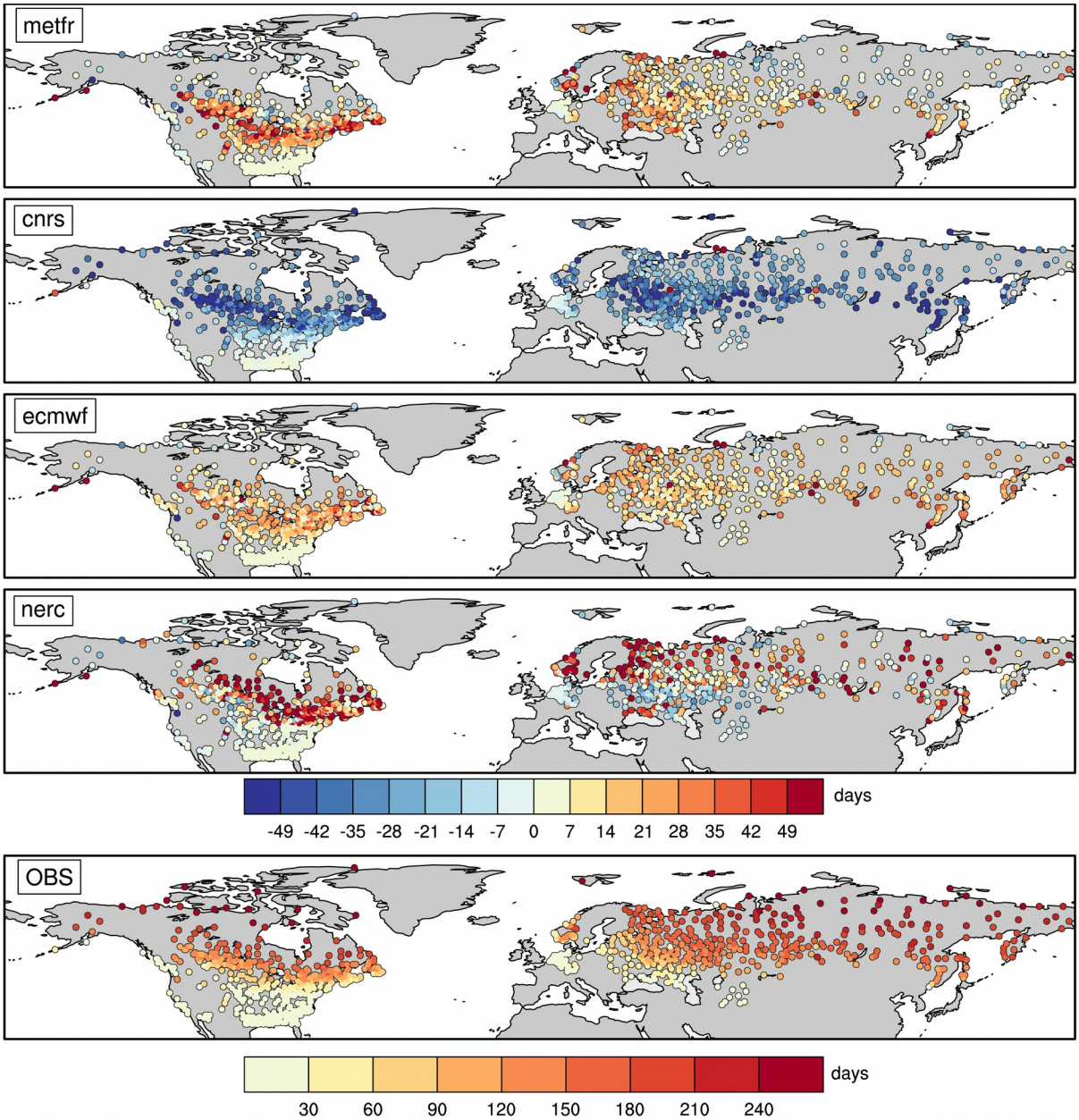

Fig. 5.2 Bias (days since 1st August) of duration of continuous snow between models and observations over the 1979-2012 period. Bottom : climatological observed values (days).

Figure 5.3 represents the first day (since 1st August) with continuous snow on the ground observed in average (bottom map). The bias (in days) of the simulated date of this first day for each model is represented by the four top maps. The biggest bias is found in NERC. METFR and ECMWF tend to simulate an earlier continuous snow period than observed, while in the CNRS experiment, the date of the first day with continuous snow occurs later.

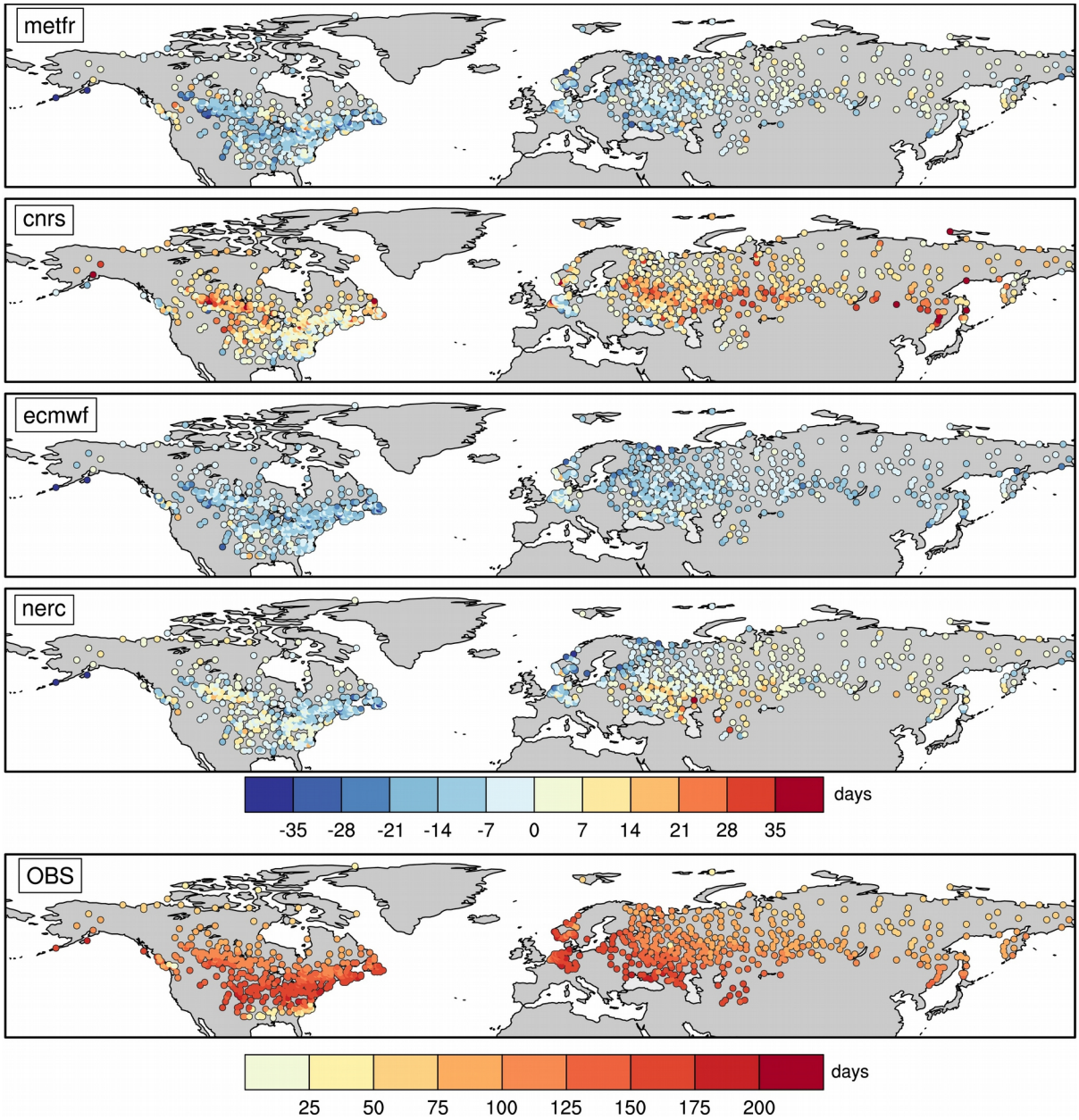

Fig. 5.3 Bias (days since 1st August) of first day of continuous snow period between models and observations over the 1979-2012 period. Bottom : climatological observed values (days).

Figure 5.4 is identical to Figure 5.3 but for the last days (since 1st August) of the continuous snow period. The bias (in days) of the simulated date of this last day for each model is represented by the the four top maps. This time, the NERC experiment presents a strong positive bias, corresponding to a later snow cover, especially over high latitudes. Bias in METFR is positive too but lower than NERC. In ECMWF, the bias is positive but relatively weak, while in the CNRS experiment, the date of the last day with continuous snow occurs much earlier.

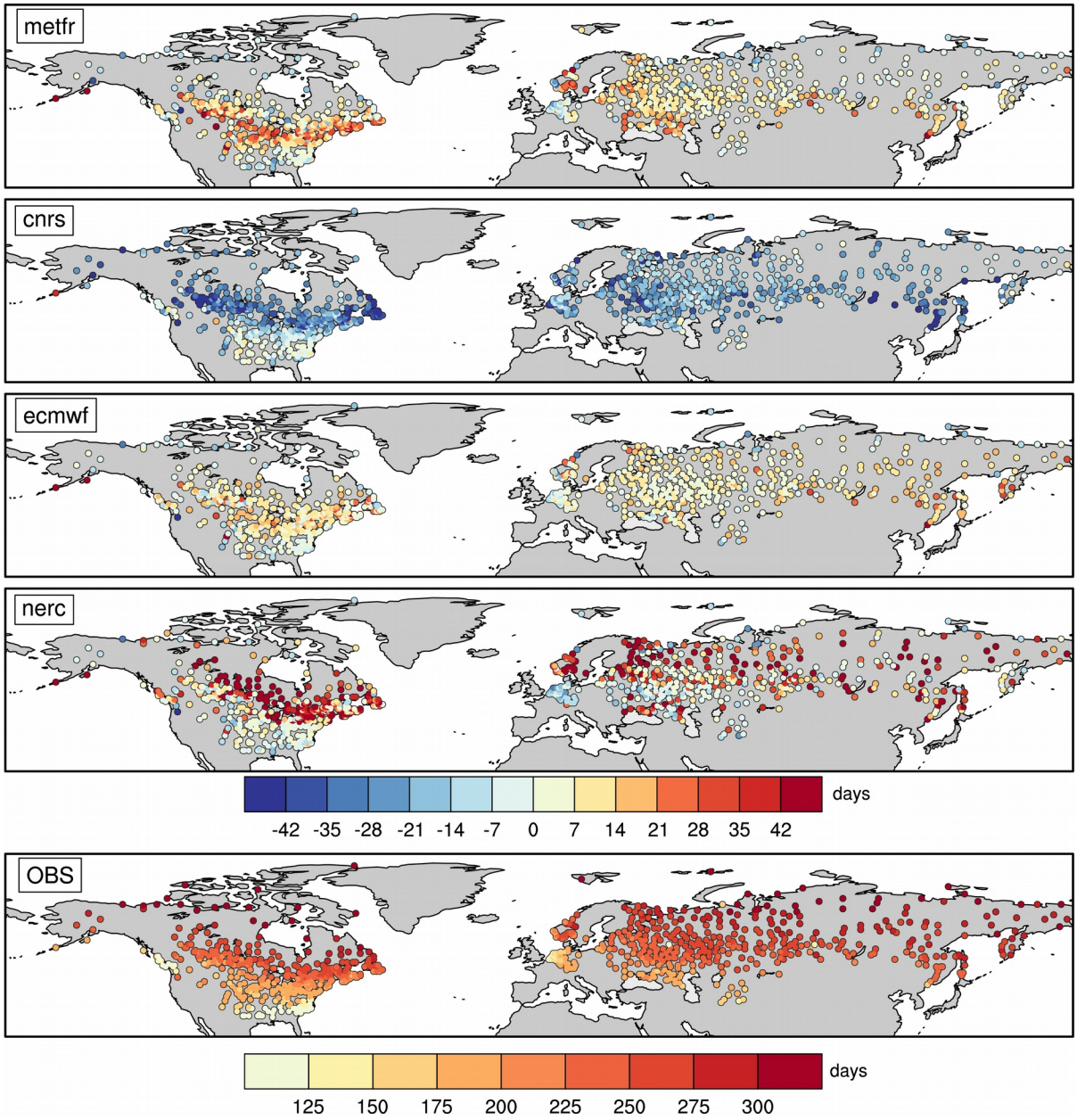

Fig. 5.4 Bias (days since 1st August) of last day of continuous snow period between models and observations over the 1979-2012 period. Bottom : climatological observed values (days).

Another way to summarise the previous maps and results, is to represent the observed and simulated annual cycle of snow depth (Figure 5.5). The same results as previously described are highlighted: two models, METFR and ECMWF simulate a snow period that is too long, and consequently, an annual mean snow depth that is higher than observed. On the contrary, NERC, but particularly CNRS show an annual mean snow depth too shallow.

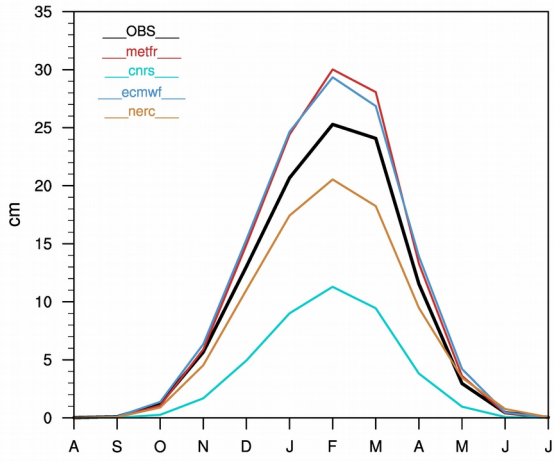

Fig. 5.5 Annual cycle for observed snow depth, and simulated snow depth over the grid points corresponding to the stations.

As shown by the previous figures, the mean state biases are evenly distributed throughout the world. We have consequently chosen to summarise the models performances in the following table (Table 5.1) which groups global bias, correlation and RMSE for the annual mean snow depth, the number of days with snow in a year, the duration and first and last day of continuous snow period.

		OBS	METFR	CNRS	ECMWF	NERC
Annual average snow depth (m)	Mean	0.069	0.082	0.028	0.083	0.058
	Bias		0.013	-0.041	0.013	-0.011
	Correlation		0.83	0.75	0.83	0.80
	RMSE		0.062	0.079	0.063	0.059
Number of days with snow per year	Mean	89.8	95.2	62.9	100.0	97.1
	Bias		5.4	-26.9	10.3	7.3
	Correlation		0.97	0.94	0.98	0.96
	RMSE		20.4	39.4	21.0	28.2
Annual duration of continous snow cover (days)	Mean	78.1	89.8	56.6	89.4	92.3
	Bias		11.7	-21.5	11.3	14.2
	Correlation		0.95	0.93	0.97	0.93
	RMSE		28.8	36.6	24.3	35.3
First day with continous snow (number of days since 1st August)	Mean	129.9	123.1	135.0	124.6	125.5
	Bias		-6.9	10.3	-7.7	-2.0
	Correlation		0.84	0.80	0.86	0.84
	RMSE		23.4	26.0	23.1	22.1
Last day with continous snow (number of days since 1st August)	Mean	231.1	239.8	217.5	235.6	251.8
	Bias		8.6	-19.7	6.3	18.1
	Correlation		0.85	0.81	0.89	0.80
	RMSE		25.7	32.2	22.5	36.4

Table 5.1 Statistical scores for the different models compared to observed snow depth stations.

Another point to consider for the validation of the simulated snow depth and not discussed previously, is the ability of the models to capture the interannual variability. In the analysis of seasonal cycle, we demonstrated that results were similar at the global and regional scale. For this analysis of the interannual variability, we have chosen to examine different regions. These regions or boxes, represented in Figure 5.7, have been defined based on the available data (see the stations localisation in the previous figures of Section 5). We have worked with three boxes with a larger number of stations : A box over Canada (CAN) with 236 stations, over Russia (RUS) with 256 stations , and over Eastern North America (AME) with 187 stations. The analysis is also performed in global mean (GLO), considering all stations.

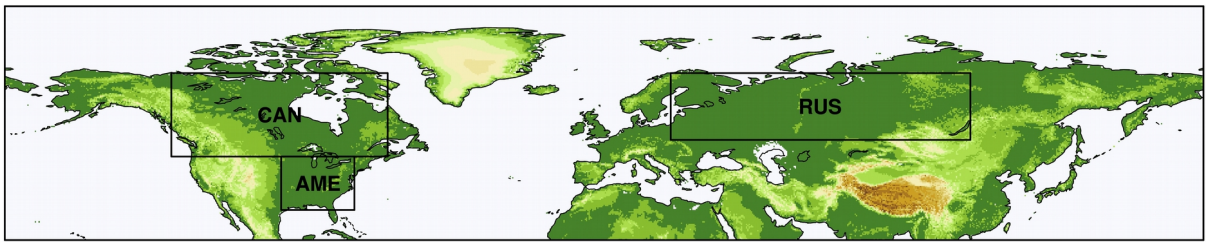

Figure 5.7 Location of the 3 geographical boxes used for the analysis of snow depth.

Figure 5.8 represents annual mean anomalies of observed and simulated snow depth over these different boxes. By representing anomalies, we avoid the mean state bias previously mentioned. It appears that the interannual variability is very well captured by the models. Correlations are between 0.78 (model METFR box RUS) and 0.95 (ECMWF box AME).

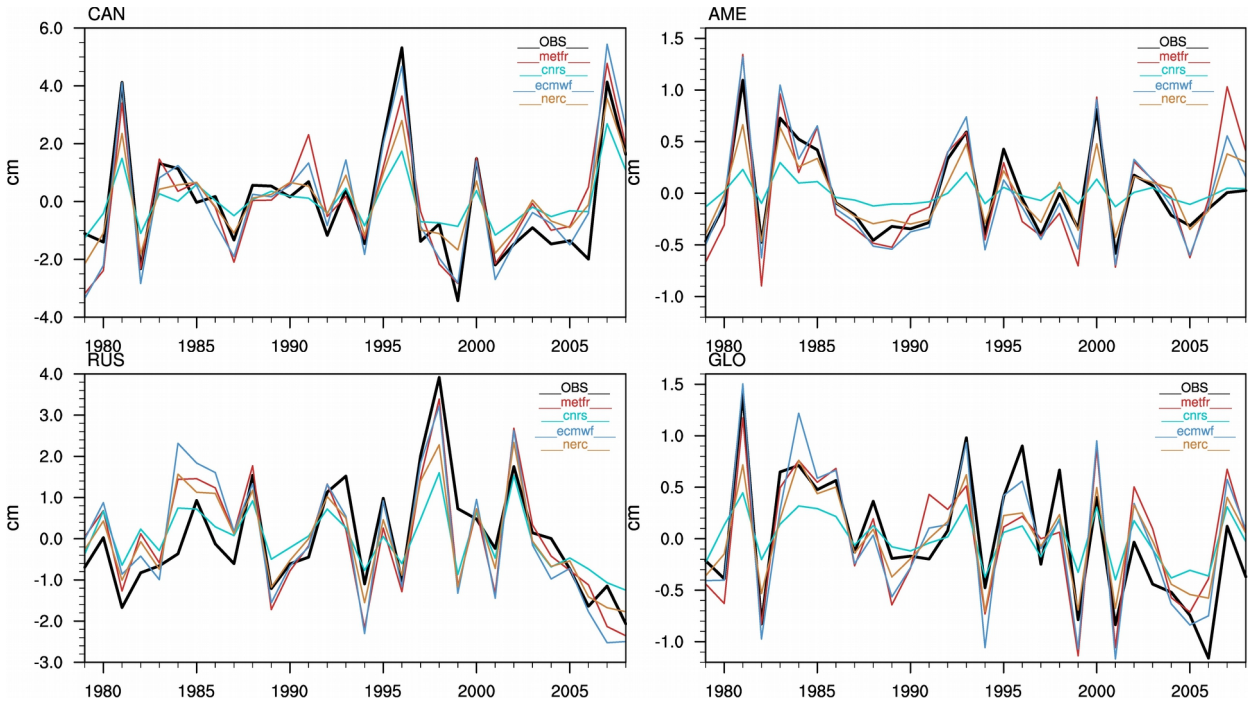

Figure 5.8 Annual mean (since 1st August) of snow depth anomalies, over the different regions: Canada (CAN), Eastern North America (AME), Russia (RUS) and global average (GLO). From 1979 to 2008. Result for 1979 corresponds to the mean from August 1979 to July 1980. In black, the observed stations. In colors, the models.

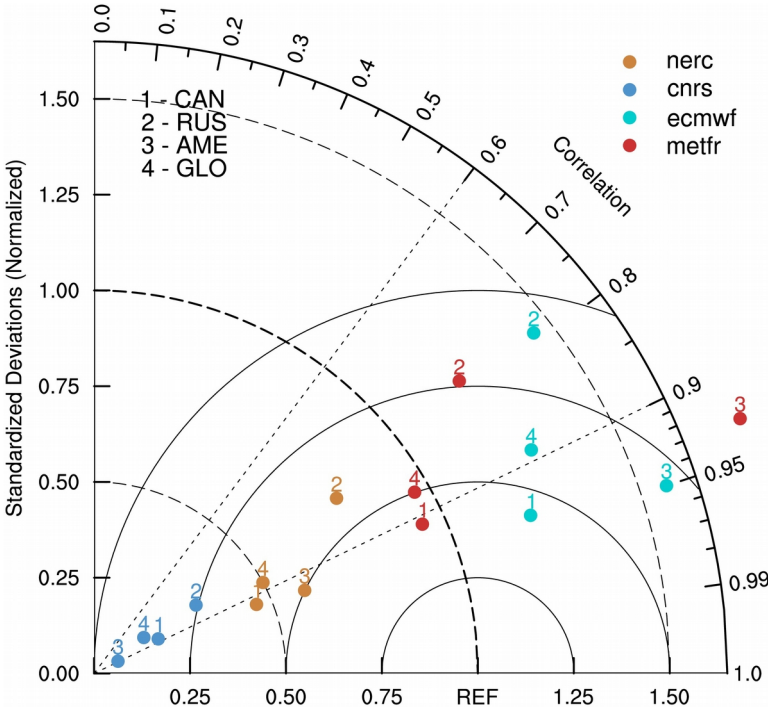

Figure 5.9 Taylor diagram for snow depth annual mean. Models are identified by colors, and numbers represent geographical boxes.

A Taylor diagram of the annual snow depth is represented Figure 5.9. Taylor diagrams are used to compare results of models to observations. A perfect model would be represented by a circle located on REF (ratio of standardized deviation and correlation equal to 1). Our four models are represented by a different color and for each models the numbers 1,2,3,4 allow us to distinguish the four geographical areas (respectively CAN,RUS, AME and GLO). Although the correlations are good, we can see important disparities in the simulated snow depth variance by the different models. METFR shows a variance that is relatively correct and close to the observed variance (ratio close to 1) except over the Eastern North America (AME). ECMWF tends to overestimate the variance. On the contrary, NERC and CNRS underestimate the variance of the annual mean of snow depth. This underestimation is especially strong in CNRS.

6 Conclusions

We have presented in this work a validation of the simulated terrestrial water storage and snow depth, in the WRR1 reanalysis.

For the TWS, the main constraint to complete this task successfully, is the difference existing between the available outputs from the models: the same variables do not exist as outputs for all models (Table 1.1), and the calculated TWS is therefore not consequently strictly comparable. However, a global analysis of the simulated TWS has been carried out to highlight principal

performances or bias of models. It has been shown that the spatial and seasonal climatologies are correctly captured (Fig. 4.1 a) and 4.2). The simulated seasonal cycle tends to be ahead of the observed one (Fig. 4.1 b) and Fig. 4.5). Biases in the representation of the variability of monthly anomalies are more dependent on the region than on the model, for instance over Canada, Sahel, Asia or Amazon, monthly correlations with the GRACE product are low for all models. This is also true for RMSE. The analysis suggests that some models such as NERC, JRC, CSIRO and CNRS present a negative bias in their estimation of the TWS variance, however it must be noted that it is precisely these models which have fewer variables for the calculation of the ΔTWS.

The validation of the snow depth was carried out 4 the models which have this variable as output: METFR, ECMWF, NERC and CNRS. A multimodel approach is difficult with so few models and has not been considered here. Biases are spatially evenly distributed. The annual mean snow depth tends to be overestimated by METFR and ECMWF, and underestimated by NERC and especially by CNRS. The seasonal cycle is in phase with the observed one, but longer for METFR, ECMWF and NERC, and shorter for CNRS. The interannual variability is generally well captured by the models, although underestimated by CNRS. CNRS model has shown significant biases for all variables (average, duration of snow period, etc...), in terms of both the mean state or the variability, and we suspect maybe a problem or bug in their simulation. As things currently stand, this simulation would be probably excluded for a multimodel study.

References

Alkama, R., B. Decharme, H. Douville, M. Becker, A. Cazenave, J. Sheffield, A. Voldoire, S. Tyteca and P. Le Moigne, 2010: Global evaluation of the ISBA-TRIP continental hydrological system. Part I: comparison to GRACE terrestrial water storage estimates and in situ river discharges. J. Hydrometeor., 11, 583-600, doi: 10.1175/2010jhm1211.1.

Brown, R.D. and R.O. Braaten, 1998: Spatial and temporal variability of Canadian monthly snow depths, 1946-1995. Atmosphere-Ocean, 36: 37-45.

Brun, E., V. Vionnet, A. Boone, B. Decharme, Y. Peings, R. Valette, F. Karbou and S. Morin, 2013 : Simulation of northern Eurasian local snow depth, mass, and density using a detailed snowpack model and meteorological reanalyses. *J. Hydrometeor.*, 14, 203-219. doi : 10.1175/JHM-D-12-012.1.

Bulygina, O.N., V.N. Razuvaev, T.M. Aleksandrova, 2014. DATA SET "SNOW COVER CHARACTERISTICS FROM RUSSIAN METEOROLOGICAL STATIONS AND FROM SOME METEOROLOGICAL STATIONS OVER THE FORMER USSR TERRITORY". Certificate of state registration №2014621201. http://meteo.ru/english/climate/descrip2.htm

Crowley, J. W., J. X. Mitrovica, R. C. Bailey, M. E. Tamisiea, and J. L. Davis, 2006: Land water storage within the Congo Basin inferred from GRACE satellite gravity data. Geophys. Res. Lett., 33, L19402, doi:10.1029/2006GL027070.

Decharme, B., R. Alkama, E. Douville, M. Becker, and A. Cazenave, 2010: Global evaluation of the ISBA-TRIP continental hydrological system. Part II: Uncertainties in river routing simulation related to flow velocity and groundwater storage. J. Hydrometeor., 11, 601–617.

Rodell, M., J. S. Famiglietti, J. Chen, S. I. Seneviratne, P. Viterbo, S. Holl, and C. R. Wilson, 2004: Basin scale estimates of evapotranspiration using GRACE and other observations. Geophys. Res. Lett., 31, L20504, doi:10.1029/2004GL020873.

Seo, K.-W., C. R. Wilson, J. S. Famiglietti, J. L. Chen, and M. Rodell, 2006: Terrestrial water mass load changes from Gravity Recovery and Climate Experiment (GRACE). Water Resour. Res., 42, W05417, doi:10.1029/2005WR004255.

Schmidt, R., and Coauthors, 2006: GRACE observations of changes in continental water storage. Global Planet. Change, 50, 112–126.

Swenson, S., J. Wahr, and P. Milly, 2003: Estimated accuracies of regional water storage variations inferred from the Gravity Recovery and Climate Experiment (GRACE). Water Resour. Res., 39, 1223, doi:10.1029/2002WR001808.

Tapley, B. D., S. Bettadpur, J. C. Ries, P. F. Thompson, and M. M. Watkins, 2004: GRACE measurements of mass variability in the Earth system. Science, 305, 503–505.

Vergnes, J.-P., and B. Decharme, 2012: A simple groundwater scheme in the TRIP river routing model: global off-line evaluation against GRACE terrestrial water storage estimates and observed river discharges. Hydrol. Earth Syst. Sci., 16, 3889-3908, doi:10.5194/hess-16-3889-2012.

Wahr, J., S. Swenson, V. Zlotnicki, and I. Velicogna, 2004: Time-variable gravity from GRACE: First results. Geophys. Res.Lett., 31, L11501, doi:10.1029/2004GL019779.

Yeh, P. J.-F., S. C. Swenson, J. S. Famiglietti, and J. Wahr, 2006: Remote sensing of groundwater storage changes in Illinois using the Gravity Recovery and Climate Experiment (GRACE). Water Resour. Res., 42, W12203, doi:10.1029/2006WR005374.