Climate Modelling User Group

Deliverable 4.1

Exploiting CCI products in MIP experiments

Centres providing input: Met Office, MPI-M, IPSL, BSC

Version	Date	Status
0.1	29 July 2019	First input from partners of activity and results
0.5	12 Sept 2019	Input from BSC, IPSL, Met Office
1.0	30 Sept 2019	Submitted to ESA
1.1	19 Dec 2019	Input from partners to address ESA RIDv1
1.2	12 Feb 2020	Revised version for submission to ESA
1.3	13 Mar 2021	Update to 1.2 with input from MPI-M
2.0	17 Sept 2021	Version 2 submitted to ESA
2.1	14 Oct 2021	Revised version addressing ESA RIDv2 submitted to ESA

CMUG CCI+ Deliverable 4.1

Exploiting CCI products in MIP experiments

Table of Contents

Report on exploiting CCI products in MIP experiments

1. Purpose and scope of this report

This document is the second report on the outputs of the CMUG Model Inter Comparison (MIP) type experiments using data products from the CCI+ ECV projects. Its purpose is to provide feedback to ESA and the CCI teams on the suitability and application of CCI climate data products in climate models. This activity has eleven experiments (CMUG WPs 4.1 to 4.11) by four CMUG partners. These are all focused studies which use CMIP6 model output for the research (as opposed to conducting new model runs). Many data products from the CCI ECVs are included, and outputs from five of the new CCI+ ECVs are used (or will be when available). An overview of the key features of the experiments is given in Table 1.

2. CMUG approach for assessing quality in CCI products

This work is concerned with exploiting CCI products in MIP experiments, with the activity in CMUG WPs 4.1 to 4.6 on statistical analyses that evaluate facets of model behaviour in representing climate. They carefully target individual elements of uncertainty derived either from the climate system (e.g.,, internal variability, system memory) or the observations (e.g.,, levels of processing or scales of averaging) and then provide a framework for combining these. There is an emphasis on characterising and understanding uncertainty in these experiments to inform the CMUG work on the ESMValTool to include uncertainty in its evaluation process for the metrics of the ECVs in these experiments. CMUG WP 4.7 addresses the important issue relevant to the component of CMIP6 focusing on decadal prediction by applying multiple CCI/CCI+ atmospheric and marine ECVs to generate an assessment of the skill in decadal forecasting systems. CMUG WPs 4.8 to 4.10 focus on the application of CCI/CCI+ terrestrial ECVs to evaluate the physical basis of representation of biophysical land surface processes and assess their simulation in earth system model components. They use data from the CMIP6 archive to understand plant climate interactions, their representation in climate models and evaluate model performance and suggest areas for future model development. WP 4.11 will build on the process analysis undertaken elsewhere in CMUG and will apply several ECVs and other datasets to identify the drivers of biases in the state of the terrestrial surface and the fluxes generated by its interaction with the atmosphere. This will provide an assessment of the value of combining multiple ECVs with other data sources to assess the quality and identify areas for improvement in the atmospheric model component of CMIP.

The uncertainty characterisation accompanying the CCI ECV datasets is examined to understand its usefulness in the modelling studies. The different types of uncertainty characterisation (grid point, bias, statistical, variance, temporal/spatial, or other) provided by the CCI ECV teams and how it meets user requirements is commented on in this report.

CMUG WP	EXPLOITING CCI PRODUCTS IN MIP EXPERIMENTS	CMUG LEAD	EXPERIMENT TYPE	CCI ECVS	OTHER ECVS
4.1	Evaluation of modelled system memory	MPI-M	Statistical analysis	Salinity, Snow, LST, SST, SI	
4.2	Evaluation of model results considering observational uncertainty	MPI-M	Statistical analysis	Salinity, Snow, LST, SST, SI	
4.3	Evaluation of model results considering the abstraction level of observational products	MPI-M	Statistical analysis	Salinity, Snow, LST, SST, SI	
4.4	Optimal spatial and temporal scales for model evaluation	MPI-M	Statistical analysis	Salinity, Snow, LST, SST, SI	
4.5	Evaluation of model results considering internal variability	MPI-M	Statistical analysis	Salinity, Snow, LST, SST, SI	
4.6	Evaluation of model results considering a combination of sources of uncertainties	MPI-M	Statistical analysis	Salinity, Snow, LST, SST, SI	
4.7	Skill assessment of the DCPP decadal predictions	BSC	Skill analysis	Sea Level, SST, Clouds	
4.8	Use LST products to develop and test simple models relating the LST versus air temperature (near surface) difference to vegetation moisture stress	Met Office	MIP process analysis	AGBiomass, LST, SM, LC	Temperature, Precipitation, FAPAR, LAI
4.9	Use CCI+ products and simple models developed in WP4.8 to evaluate performance of LST versus air temp, using multiple land surface and ES models	Met Office	MIP process analysis	LST, AGBiomass, LC/HRLC	Temperature
4.10	Comparison of CCI data in vegetation study with other satellite data and LS models	Met Office	MIP process analysis	AGBiomass, LST, SM, LC	Temperature, Precipitation, FAPAR, LAI
4.11	Land-surface interaction related biases in AMIP	IPSL	MIP process analysis	LST , Snow, SM	Air temp, turb. fluxes (Jung, Gleam,) meteo analysis, MODIS data, CERES rad. fluxes, SM (SMOS, Gleam)

Table 1: Main features of the work on exploiting CCI products in MIP experiments.

3. Links between Task 4 and the CMIP projects

The results are relevant to the CMIP6 endorsed MIP projects that are working in a similar research area to CMUG WP4. CMIP is part of WCRP (World Climate Research Programme) which has proposed areas for emphasis in climate research called the 'grand challenges'[1], which the MIPs are helping to address. There are currently 23 CMIP6 endorsed MIP projects[2] (plus 17 related or supporting MIP type projects) which cover a wide range of Earth system processes and modelling activities. The CMUG partners working on this Task are engaging with relevant CMIP projects and exchanging results and information about their respective research. There are CMUG partners are currently involved in all CMIP projects as summarized in Table 2.

CMUG PARTNER	MIP PROJECTS
Met Office	All MIP projects, either directly or through collaborative research with the UK institutes using the Met Office climate model The Met Office leads HighResMIP. A Met Office researcher is a panel member for CMIP6
DLR	Veronika Eyring is a panel member for CMIP6 MIPs relevant to atmospheric processes and chemistry
IPSL	LS3MIP (Land Surface, Snow and Soil Moisture Model Intercomparison Project) - the results will be valuable for the work proposed in CMUG. SPMIP for Soil Parameter MIP - the results will be particularly valuable for the work proposed in CMUG. AMIP HiResMIP (an AMIP at higher resolution)
BSC	ScenarioMIP (5x SSP2-4.5 scenario runs) DCPP VolMIP (volcpinatubo-full and volc-long-eq) HiResMIP PRIMAVERA: spinup, hist-1950, control-1950 and highres-future) AerChemMIP (piClim-2xdust) C4MIP OMIP DECK
MPI-M	Dirk Notz is co-chair of SIMIP Researchers at MPI-M are involved in virtually all MIPs and will provide respective model output from specific simulations.
Météo France	AERCHEMMIP CFMIP DAMIP DCPP FAFMIP LS3MIP RFMIP ScenarioMIP CORDEX Plus an involvement with many others
SMHI	CMIP HighRESMIP

Table 2: Summary of CMUG involvement with CMIP projects.

[1] https://www.wcrp-climate.org/grand-challenges/grand-challenges-overview
[2] https://www.wcrp-climate.org/modelling-wgcm-mip-catalogue/modelling-wgcm-cmip6-endorsed-mips

4. CMUG MIP experiments with CCI products

4.1 Evaluation of modelled system memory

Lead partner: MPI-M

Author: Andreas Wernecke

Aim

The aim of this research is to develop and apply a framework that allows evaluation of the simulated memory (temporal correlation) of ECVs in a model-evaluation processing chain. It will address the following scientific question: How can we evaluate the memory of climate variables as simulated by large-scale model simulations?

Summary of Work and Results

Work on this experiment has so far focused on the Sea Surface Salinity ECV (SSS) but the general workflow is adjustable to other ECVs. The temporal autocorrelation, or memory, is an essential property of all ECVs. It describes the ability of the earth system to maintain a quantity despite climate variability. The memory is also closely related to the predictability of a variable and the timeframe for which data assimilation into prediction models has the potential to be beneficial. However, here we do not investigate the role of model memory in the context of (e.g., seasonal-) prediction models but for the evaluation of climate models in general. The memory of a system variable is the result of the sum of all relevant physical processes, acting on their respective time scales. A disagreement of modelled and real memory indicates either that the relevance of processes is falsely interpreted (including potentially neglecting a process completely) or that processes are misrepresented in the model so that the respective relevant time scales are wrong. Observational uncertainties can also distort the image of the real system memory where, for example, sensor white noise would reduce the observed memory.

Three statistics are used here, which are quickly introduced in the following.

The Anomaly Correlation Coefficient (ACC) is frequently used as fully localized measure of the correlation between a seasonal forecast (v) and observations (o). The ACC is the Pearson correlation coefficient for a given location and month of the year, calculated over a range of years. For the memory we treat the SSS of month x as forecast for a following month (x+lag). For example, the ACC can be a measure of how strongly a positive SSS anomaly in, say, January is informative for the SSS anomaly in March (lag=2 month), at any given location.

The lagged pattern correlation on the other hand is defined as the Pearson correlation between two time slices, calculated across all locations. The pattern correlation between January and March can therefore have different values for each year, which we average using a Fisher-Z transformation. The pattern correlation is a global statistic describing regional memory; how long does a pattern (fingerprint) of regionally high/low anomalies persist? Limitations are the disregard towards biases (of limited concern here since the seasonal cycle/trends are not the focus area) and amplitude of the pattern. For example, if there is a spatial pattern of high/low anomalies in January which diminish homogeneously (anomalies become smaller with time but maintain their relative spatial distribution) the pattern correlation would still attest full/perfect memory.

Lagged Mean Squared Differences (MSD) (also called mean squared error) reflect a combination of change in amplitude and change in location. The MSD is easily converted into a skill score by S_MSD=1-MSD/MSD_REF, setting one to a perfect value (no differences) and zero to an MSD equal to a reference MSD. Typically reference values are based on an earlier MSD or a climatology (Section 8.3.3 in 'Statistical Methods in the Atmospheric Sciences'; Wilks, 2019). Here we use the climatology as reference so that S_MSD=0 corresponds to a lag time where the initial SSS anomaly is just as good a predictor for a later time as the climatology (i.e. zero for anomalies). Note that for lag times larger than the temporal correlation length scale (memory), the climatology is the best a-priory predictor of the SSS state, meaning that negative S_MSD are to be expected.

Here we investigate SSS memory in the CCI+ SSS product, ORAS5 reanalysis and the MPI-ESM grand ensemble (MPI-GE). We use two periods for model-to-observation comparisons which are 1979-2005 (ORAS5 and MPI-GE historical runs) and 2010-2019 (CCI and MPI-GE RCP4.5 runs). In all cases we first derive the (linearly) detrended anomalies (the respective climatologies are based on the same periods as mentioned before) and bring the MPI-GE data onto the observational (EASE-2) projection. The MPI-GE sea surface salinity extends underneath sea ice, where the view for CCI satellite observations is blocked. We use only locations for which valid SSS observations are available throughout the whole data period and use the same mask for MPI-GE data. To minimize the influence of sea ice further we limit the study area to 65° S to 65° N.

Figure 4.1.1 illustrates crucial steps towards the SSS memory analysis. The CCI+ SSS product and MPI-ESM data are brought to the same grid and masks (location of valid estimates) are synchronized (top row of Figure 4.1.1). SSS anomalies are derived year-round by subtracting the seasonal cycle and linear trends (examples for January and April 2019 shown in the second and third row of Figure 4.1.1). The anomalies (and with that the absolute SSS values) cannot be expected to agree between the data sets since they represent internal variability of the (modelled and observed) system. Again, the model runs used here are climate projections and are not intended to forecast the one realization of internal variability which the real world is taking but instead to represent plausible (alternative) realizations with realistic magnitude, spatial and temporal characteristics. The memory, represented by the ACC in the bottom row of Figure 4.1.1, is one of those characteristics which would ideally be consistent between the data sets. Overall, the ACC is considerably smaller in the CCI+ product than in the MPI-ESM and fewer locations have significant correlation. That being said, some regional similarities do exist, for example in the tropical pacific with bands of increased memory north and south of the equator as well as the north Atlantic between 10° N and 30° N and around Australia and Maritime Southeast Asia.

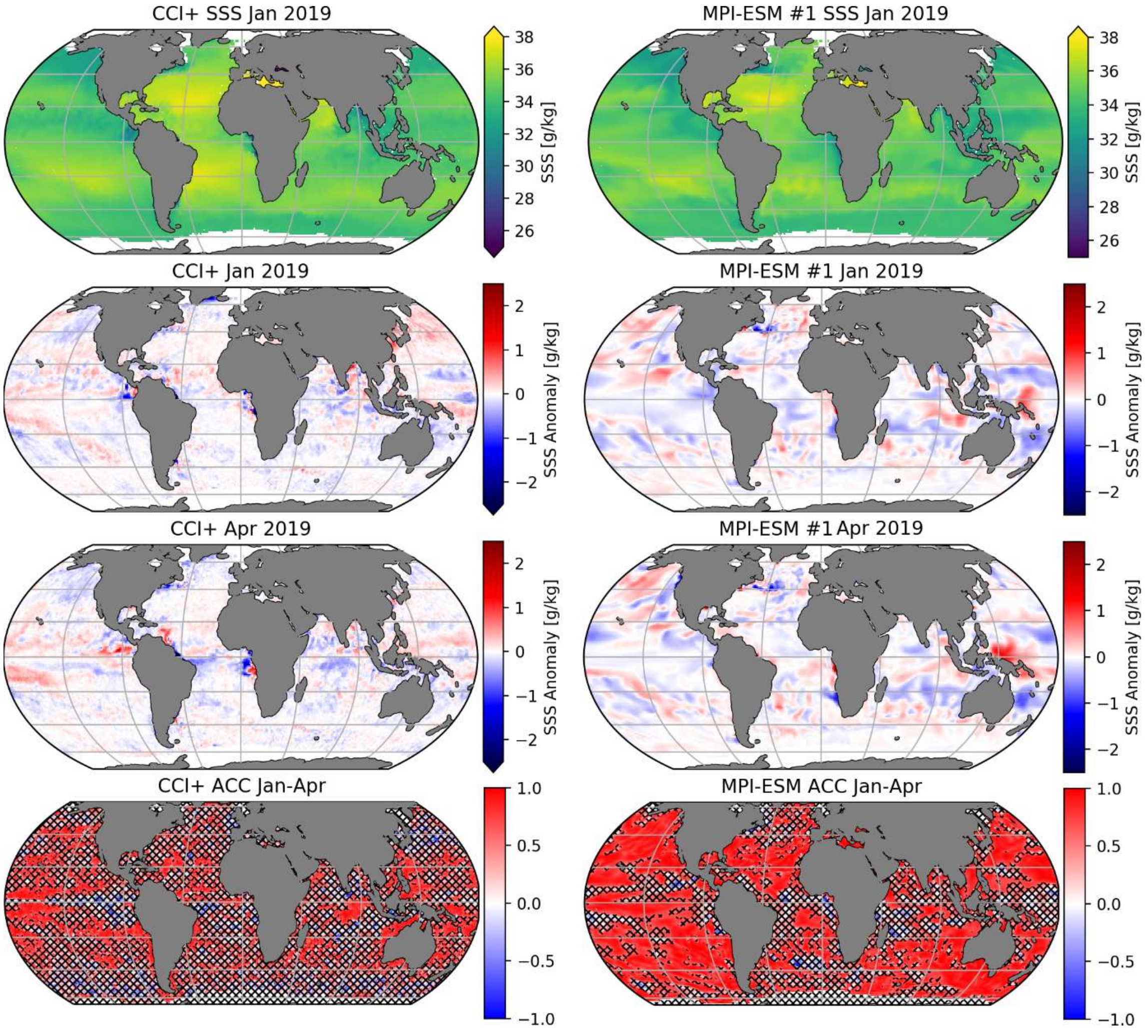


***Figure 4.1.1**: CCI+ satellite observations (left) and MPI-ESM grand ensemble member #1 (right) January 2019 Sea Surface Salinity (SSS) (top) and SSS anomalies for January and April 2019 (second and third row respectively) and Anomaly Correlation Coefficients (ACC, bottom row) between January and April SSS based on 2010 to 2019. Locations with ACC p-value below 0.05 (failed significance test) are hatched.*

The local memory, as approached above by the ACC, can give valuable information of regional model to observation agreement. Relevant processes, leading to agreement or disagreement between the data sets, will however change throughout the year and act on a range of timescales, making a systematic investigation challenging (note that we show only the ACC between January and April as examples). The main global pattern appears to be that the observed memory is shorter than the MPI-ESM memory making local interpretations cumbersome. To test this hypothesis, we use the global anomaly pattern correlation.

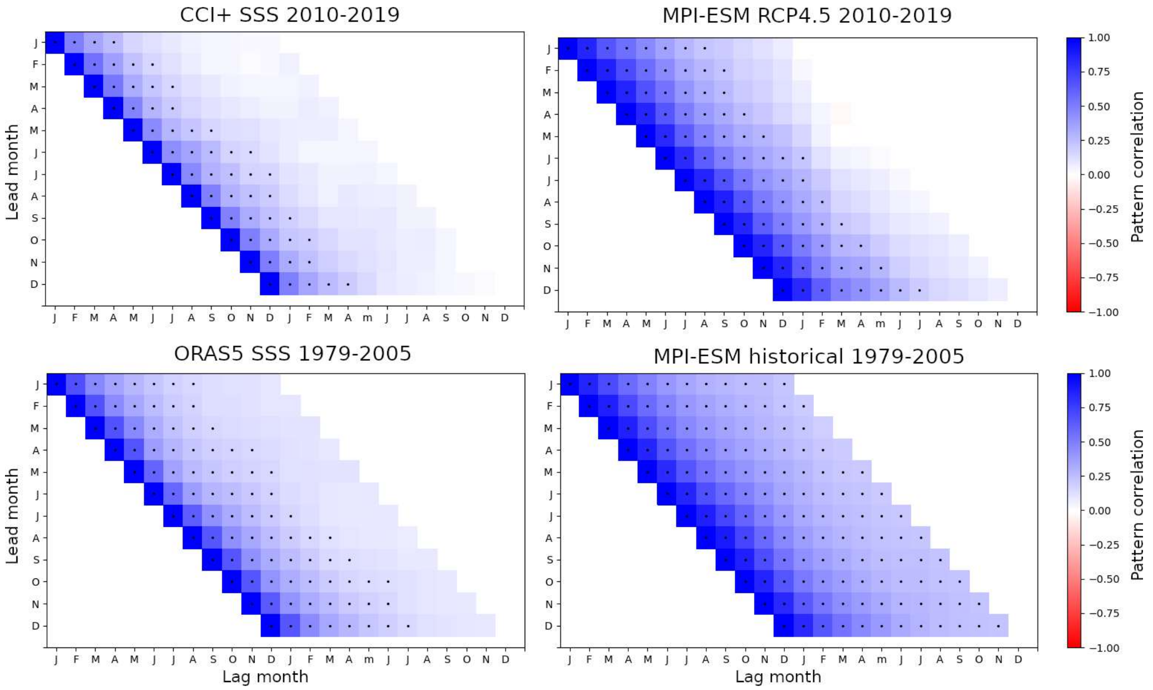


__Figure 4.1.2__: Global SSS anomaly pattern correlation from MPI-ESM (right) and observations (left), namely the CCI+ SSS product (top) and ORAS5 ocean reanalysis (bottom). Note that the MPI-ESM data are confined to the same time periods as the respective observations and that we use historical forcing experiment before 2005 and RCP4.5 experiment past 2005. Dots indicate significance.

We calculate the lagged pattern correlation between each month of the year (lead month) and each month of the full succeeding year (lag month). The lag period goes therefore from zero to eleven month (x-axis of Figure 4.1.2) where zero lag time corresponds to a perfect correlation of one. In general, the memory characteristics are discussed in terms of persistence (the initial short-term drop in correlation), long-term memory and potential reemergence of correlations throughout the year. However, the results shown in Figure 4.1.2 do not show noteworthy variations throughout the year or any features but a monotonic drop in pattern correlation. The typical time scale of these drops differs however by data set. The CCI+ product shows the shortest memory of only about three months; followed by ORAS5 reanalysis with about five months; the ten year MPI-ESM RCP4.5 sub-period of about six to seven months; and the longest memory of the 27 year MPI-ESM *historical* sub-period of more than 12 months (Figure 4.1.2).

While noise in the satellite data could result in an underestimation of the system memory, we do not expect the ORAS5 reanalysis data to be particularly noisy. The consistently longer memory in model runs compared to both types of observations therefore suggests that the MPI-ESM grand ensemble simulations have an unrealistically long modelled system memory. The temporal evolution of these model simulations is therefore apparently too smooth on short (seasonal to yearly) time-scales. As mentioned before, the MPI-ESM RCP4.5 memory for the

CCI+ period (2010-2019) is significantly shorter than for the *historical* simulations from 1979-2005. This is fully consistent with shorter memory in the CCI+ data compared to ORAS5 data which is therefore no indication for a substantial influence of noise on the CCI+ data.

The pattern correlation does not cover changes in the amplitude of the anomalies but just the relative spatial high/low anomaly distribution, which is why we complement the pattern correlation by an analysis of the MSD skill score.

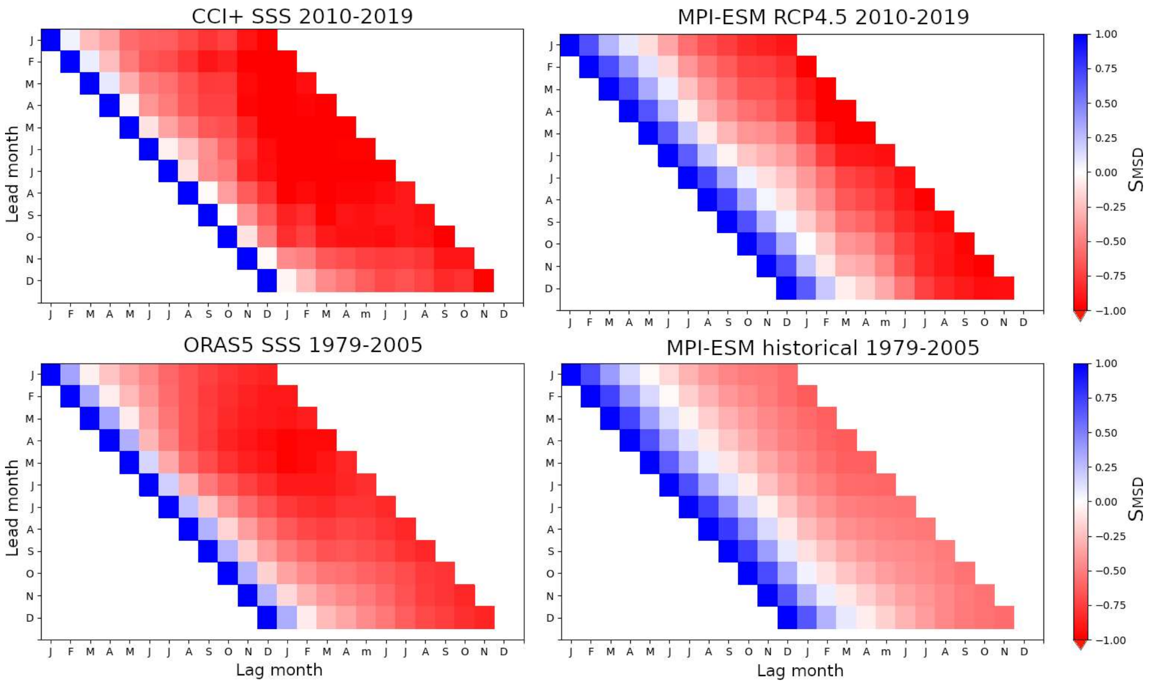


Figure 4.1.3: *As Figure 4.1.2 but showing the Mean Squared Differences Skill Score (S_MSD) instead of the pattern correlation.*

The lagged MSD Skill Score (Figure 4.1.3) shows the same behaviour as the pattern correlation with the only difference that small values are reached after shorter periods which can be easily explained by the differences in the statistics. Both show a value of one for optimal agreement but while zero in the pattern correlation indicates no correlation between the two anomaly fields, an MSD Skill of zero only indicates that the MSD is as small as the climatology value. The order of length of memory by data set is the same for both statistics.

Since the memory has no clear dependency on the time of the year, as can be seen in Figure 4.1.3, we illustrate the decline of S_MSD as function of lag time and combine model and observational estimates in Figure 4.1.4. We further define an estimate for the memory as the first crossing of S_MSD with zero, i.e. the time for which the lead month can be considered useful, i.e. better than climatology, predictor for the lag month. Note that the absolute value of this definition of memory is strongly dependent on the statistic used (compare Figures 4.1.2 and 4.1.3) but is nevertheless useful for model to observation comparisons. It can be seen that MPI-ESM simulation can largely be separated from the observations for both data sets without consideration of seasons or any temporal averaging (Figure 4.1.4).

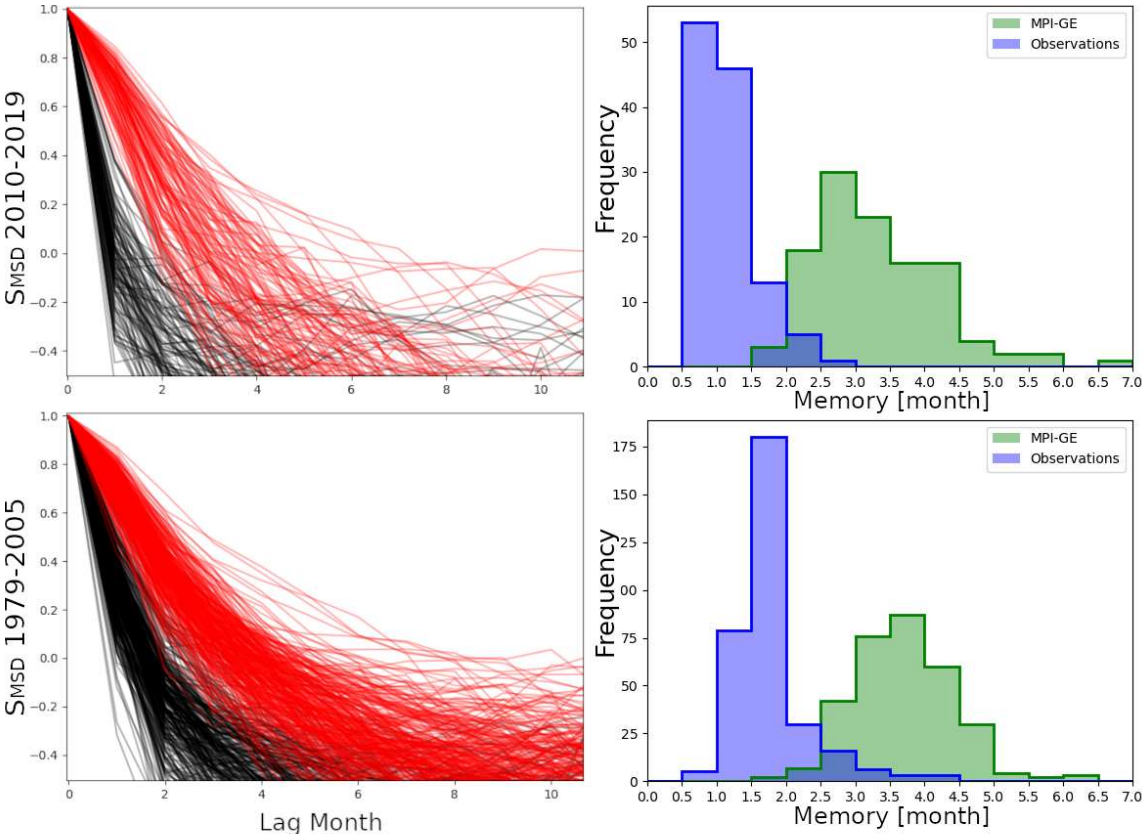


Figure 4.1.4*: Decline of S_MSD with lag time for observations (black) and MPI-ESM (red) (left) where there is one line for each month of the time series, representing the decline in skill with the following month. The frequency distributions of the corresponding first crossing of S_MSD with zero using linear interpolation ('memory') is shown on the right. The CCI+ data and period are used for the top row and ORAS5 data and period for the bottom row.*

Lastly, we investigate the potential of a latitudinal dependency of the memory, inspired by the findings from the ACC above. For this reason, we derive the memory (as defined above) on latitudinal bands of 15° for each month of the year. Besides the now well-established difference in absolute memory we see good agreement in the latitudinal-temporal development. The memory is largest at around +/-30° N and towards 60° N with the shortest memory near the equator (Figure 4.1.5). Also, the temporal development shows many similarities with a prolongation of the memory in the first half of the year (January to June) around 30° N and in the second half of the year (July to December) at around 30° S. Small differences between observations and model are however noticeable; the tropical minimum in the MPI-ESM data lies between approximately 10° S (around June/July) to 10° N (December/January) while those points are about 5° to 10° further north and about two month earlier in both observational data sets.

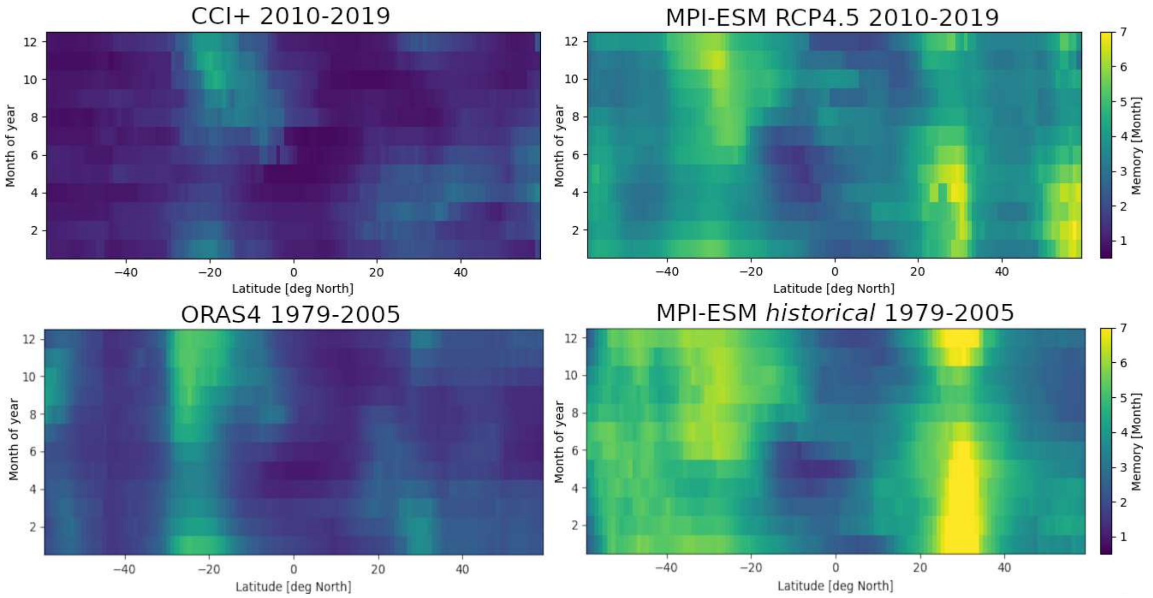


Figure 4.1.5: The memory (defined here as the first zero crossing of the MSD Skill Score) by latitude (15° bands centreed at the latitudes shown) and lead month, averaged over the years and based on the data set, noted above each panel.

Publications

None so far, but we plan to describe related concrete plans in the next version of this report.

Interactions with the ECVs used in this experiment

Interactions between the CMUG and ECV projects for work on this WP in particular happened though an email exchange with the CCI+ SSS science lead where the influence of the sensor penetration depth on the characteristic depth of the surface water layer have been discussed. We concluded that under most circumstances (all but strong rain events) SSS satellite observations are representative for the surface mixed layer, which allows a one to one comparison with modelled SSS memory. The quarterly CSWG and the Integration meetings allowed for additional interactions, including with the SSS team.

Consistency between data products

So far we did not identify any inconsistencies of concern with regard to the system memory (in addition to those discussed above). There appears to be a bias in the global mean SSS between the model and observations of about 0.2 g/kg to 0.25 g/kg, which appears to be larger in the southern hemisphere than in the northern hemisphere. These biases towards MPI-ESM data are, however, consistent between reanalysis and CCI+ observations, suggesting that the model has a negative bias towards the real state.

Recommendations to the CCI ECV teams

To be completed in next version of this report.

4.2 Evaluation of model results considering observational uncertainty

Lead partner: MPI-M

Author: Andreas Wernecke

Aim

The aim of this research is to develop and apply a framework that allows one to include observational uncertainty information into a model-evaluation processing chain. It will address the following scientific question: How can we take observational uncertainty into account when evaluating large-scale model simulations?

Summary of Work and Results

So far, work in this area has focused on the sea ice ECV. When the sea ice ECV is used for model evaluation, this is in most cases done in terms of the Sea Ice Area (SIA), Sea Ice Extent (SIE) or Sea Ice Volume (SIV). Here we focus on the SIA due to known limitations of the SIE (such as resolution dependency) and it being used much more frequently than the SIV. The SIA is calculated as the Sea Ice Concentration (SIC) multiplied by corresponding pixel size, summed up over the whole hemisphere. Sometimes the difference in SIA from a few SIC products is used as a rough estimate of the SIA but with ongoing progress in SIC uncertainty quantification, an accompanying single product SIA uncertainty estimate seems overdue. The challenge is to convert local SIC uncertainty to a combined SIA uncertainty for which it is necessary to take into account the spatial covariance structure. The importance of the correlation structure for the SIA uncertainty, and with that for model evaluations, becomes clear when considering the two extremes: All SIC pixels could be considered statistically independent which would in practice result in SIA standard deviation of order 10 000 km². The other extreme is to consider all local uncertainties throughout the hemisphere as fully correlated which would increase the SIA uncertainty in practice to the order of 1 000 000 km² (based on 50 km resolution CCI+ SIC data).

Work on this WP started with the theoretical development of a spatial covariance model which combines expected correlation signatures (based on our understanding of the error sources following discussions with the CCI+ SIC team; Thomas Lavergne, Met Norway). The main challenge is to quantify covariance model parameters (of our new model or in fact any covariance model). Most prominently this is the spatial de-correlation length scale, i.e. the characteristic spatial distance at which errors in the SIC product are largely independent. We address this challenge by three different approaches, as described below.

The covariance model

The algorithmic and smearing uncertainties (as provided by the CCI+ product) are assumed to be independent, each with their own correlation matrix.

The algorithm uncertainty is expected to be largely driven by tie-point and methodological uncertainties and only to a smaller (here neglected) extent from local measurement uncertainties. Since small SIC values will be predominantly impacted by the ocean tie-point and high SIC values predominantly impacted by the 100% sea ice tie-point, we base the (hemisphere wide) algorithmic correlation solely on differences in the sea ice concentration, not on the physical distance between measurements. The following error correlation function $(c_a(x_i,x_j))$ for the algorithmic uncertainty between measurements at locations x_i and x_j, fulfills these criteria:

$$c_a(x_i,x_j) = \exp\left[\frac{-\delta_{SIC}^2}{l_{SIC}^2}\right]$$

where l_{SIC} is a scaling parameter and δ_{SIC} is the absolute difference in SIC (in percent) at the locations x_i and x_j.

The smearing uncertainty represents a range of influences on the satellite measurements related to the different footprint sizes and spillover effects from outside of the theoretical footprints. Its correlation structure is hence more complex and should fulfill the following considerations:

- The correlation should diminish with distance between locations
- Uncertainties for similar SIC values are more likely to be subject to coherent errors than across SIC gradients
- The land spillover effect near coasts is expected to cause correlated errors.

The following error correlation function for the smearing uncertainty $(c_s(x_i,x_j))$, between measurements at locations x_i and x_j, fulfills these criteria:

$$c_s(x_i,x_j) = exp\left[\frac{-\delta_x^2}{(l_{x0} + l_{xSIC}(1-\delta_{SIC}/100\%) + l_{xL}r_L)^2}\right]$$

with l_{x0}, l_{xSIC} and l_{xL} being components of the characteristic correlation length scale, δ_x being the distance between x_i and x_j, δ_{SIC} as defined before and r_L a factor representing the combined proximity to the land by:

$$r_L = \frac{l_L^2}{\delta_{L,i}\delta_{L,j}}$$

where l_L is a typical length scale for the impact of land on the correlation and $\delta_{L,i}$ being the shortest distance to land (as defined by the SIC land mask) of x_i ($\delta_{L,j}$ is defined accordingly). To restrict the maximal impact of the land spillover on the correlation we set the maximum of r_L to one. Note that l_{xL} and l_L are separate parameters, the first representing the maximal additional correlation length scale in $c_s(x_i, x_j)$ due to land influence (which can be understood as a distance the land influence is able to carry the uncertainty correlation) and the latter (l_L) representing the typical distance away from the coast which is impacted.

In summary, we defined correlation functions that match our basic expectations. It is not encompassing anti-correlations which is in line with our error characteristic expectations. The smearing error correlations diminish with distance between two measurements with the typical correlation length-scale (at which the correlation has fallen to about 0.37) between l_{x0} and $l_{x0}+l_{xSIC}+l_{xL}$, depending on the sea ice concentration (longer with similar concentration values) and proximity to land (longer close to the coast). There are five free parameters (algorithmic and smearing uncertainty combined) which need to be set $(l_{x0}, l_{xSIC}, l_{xL}, l_L, l_{SIC})$. This is not to say that these are the only or best functional forms to represent the error correlation structure, it is one representation in line with our expectations. More research is needed to test these expectations, test other functional relationships and constrain the free parameters. The work described in the following is an early attempt to deepen our understanding in this regard.

Investigation of the CCI+ SIC error correlation

First, we derive spatial SIC correlations and SIC uncertainty correlations from repeated measurements. The rationale behind the use of the spatial SIC correlation as proxy for the SIC error correlation is that for a constant real SIC, the changes in SIC measurements would represent errors of repeated measurements. The SIC uncertainty correlation is related to the SIC error correlation by the idea that any process causing an increased uncertainty over a certain spatial footprint is more likely to cause the corresponding errors to be correlated as well. To be clear about the difference between errors and uncertainties: The uncertainty is a measure of the width of the distribution of a random variable (here the CCI SIC at a given time and location), the error is the SIC difference between a specific measurement (e.g., the SIC product value which is the centre of the uncertainty distribution) and the real value. If the uncertainty estimates are good, the error distribution will be consistent with the uncertainty estimates. For example, if two locations have highly correlated uncertainties it means that if one has relatively wide probability distribution, it is very likely that the other one has a relatively wide probability distribution as well. It does *per se* not mean that an e.g., overestimated SIC measurement at one location makes it more likely that the measurement at the other location is overestimated as well.

For Figure 4.2.1 (right) we derive the SIC correlation structure from the CCI+ SIC measurements from February 21st, 2016 and use the provided uncertainties (algorithmic and smearing) from Feb. 21, 2016 to receive a combined covariance estimate. This is compared (Figure 4.2.1, left) with the covariance based on our correlation model with selected parameters (here: $l_{x0}= 100\,km, l_{xSIC} 300\,km, l_{xL}= 100\,km, l_L= 100\,km, l_{SIC}= 20\%$). The model has the advantage that, once the parameters are derived, it is available for every time and place and adapts to changes in the ice cover dynamically. While this is just one example, it shows that sample covariance structures can show complex patterns and that the correlation model developed here is capable of representing such structures reasonably well. Note that this approach is challenging to evaluate systematically (Figure 4.2.1 shows the covariance for one day and relative to one location).

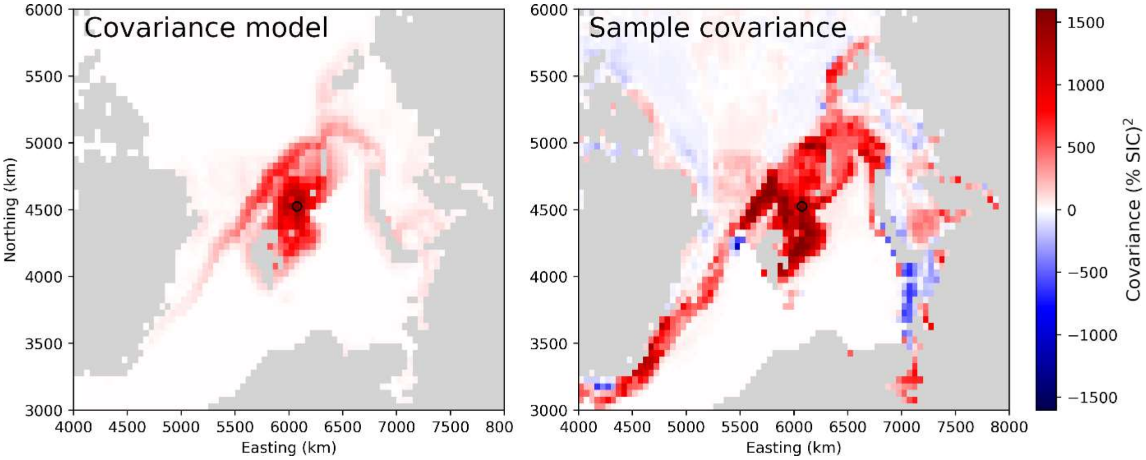


Figure 4.2.1: *Spatial covariance for selected location northeast of Svalbard (black circle) from the developed covariance model on the SIC product from Feb. 21, 2016 (left) and the sample SIC covariance with correlation pattern based on the years 2007 to 2016 for the same day (right).*

For comparison and generalization of the previous finding we now use correlation length scales from the CCI SIC validation and inter-comparison report (PVIR) (https://climate.esa.int/documents/76/Sea_Ice_Thickness_Product_Validation_and_Intercomparison_Report_1.1.pdf). These global SIC correlation length-scale estimates, kindly provided by Stefan Kern, represent the approximate circular radius of correlation in the SIC product and SIC uncertainty product. They are therefore not suited to directly constraining the parameters of our error correlation model, particularly not the non-circular components.

In the PVIR, MODIS data are used to identify regions with >=90% (labeled ’Pack Ice’) and 0% SIC (labeled ’Ocean’) and 31 day periods of the CCI product derivation from these values is used to calculate a sample correlation. The availability of MODIS SIC estimates is limited by clouds, so that this analysis is applied to windows of opportunity and represents only errors for cloud free conditions. The correlation length-scale are calculated by defining rings around the currently investigated cell and fitting an exponentially decaying function to the average correlation within each ring. This is repeated for each cell and day. The reported correlation length hence corresponds to the distance at which the correlation towards the centre cell has dropped to approximately 37%

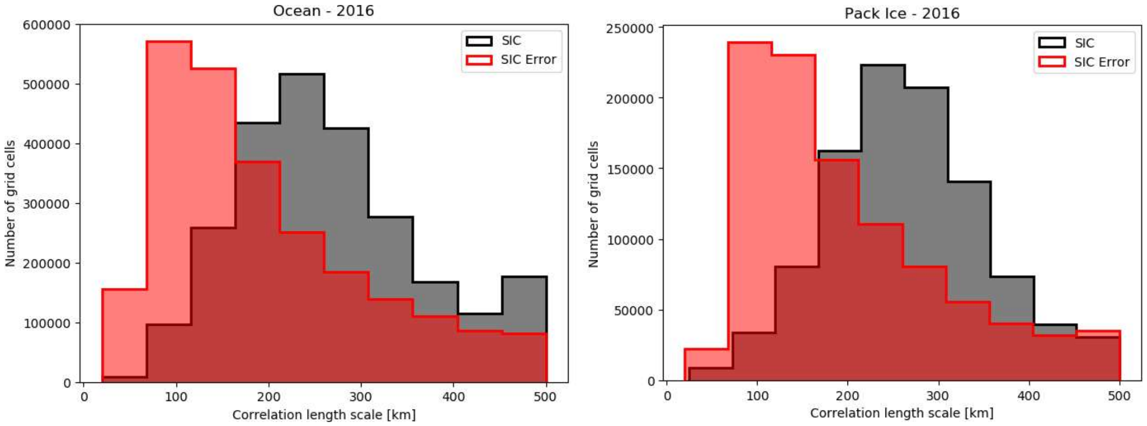
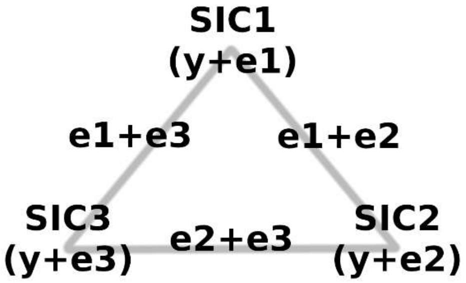


***Figure 4.2.2**: Frequency distributions of 2016 correlation length of the SIC and SIC uncertainty variables for preclassified open water locations (left) and pack ice locations (right). Based on calculations done for the PVIR.*

In Figure 4.2.2 we see the distribution of correlation length scales for the whole Arctic basin within year 2016. Other years and quarterly assessments show very similar results (not shown). The differences in SIC error correlation length scales between pack ice and ocean conditions are very small (compare left and right of Figure 4.2.2). This is a promising result since it means that there are no indications for a dependency on the correlation length on the SIC values or between typical pack ice and ocean regions. This provides no information about a potential dependency on SIC gradient. The SIC uncertainty product has shorter correlation scales than the SIC product (red vs. black histograms). The range of values is mostly between one hundred to a few hundred kilometres.

Lastly, we approach the error correlation by triangulation of independent satellite SIC products. This approach avoids any assumptions about the real state of the SIC (y) since it is the same for all SIC products and cancels out when basing the calculations solely on the differences in SIC products. Figure 4.2.3 illustrate this approach and provides the derived equations which assume that the unbiased SIC product errors (e1 to e3) are independent between the products. This assumption might not be justified, considering similarities in used satellite sensors, frequency bands and retrieval approaches.

For Figure 4.2.4 we use daily SIC fields from all days in February (excluding the 29th) for the years of 2003 to 2017 (excluding 2012 and parts of Feb. 2016 due to missing data). It is based on the CCI+ SIC, NASA team algorithm (as provided by NSIDC, doi: https://doi.org/10.7265/N59P2ZTG) and SSMI/ASI algorithm (as provided by the ICDC https://icdc.cen.uni-hamburg.de/seaiceconcentration-asi-ssmi.html). Here we can combine several years of February data due to reduced dependence of the real state of the sea ice for this approach.



$$cov(e1) \approx \frac{1}{2}(cov(Z_{12}) - cov(Z_{23}) + cov(Z_{13}))$$

$$cov(e2) \approx \frac{1}{2}(cov(Z_{12}) + cov(Z_{23}) - cov(Z_{13}))$$

$$cov(e3) \approx \frac{1}{2}(- cov(Z_{12}) + cov(Z_{23}) + cov(Z_{13}))$$

Figure 4.2.3*: Schematic of SIC error triangulation and corresponding derived equations with y representing the real SIC, ei being the error of SIC product i and Z_ij being the difference in SIC product i and j.*

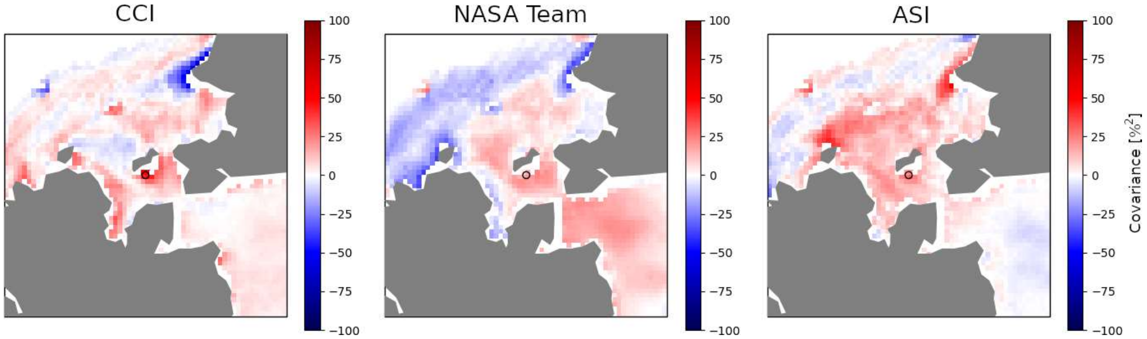


Figure 4.2.4*: Covariance estimates (equations given in Figure 4.2.3) for February (as example) of three SIC products (left: CCI+, centre: NASA team, right: SSMI ASI) relative to the location marked by a black circle, approx. 200 km south of the Bering strait and just north of St. Lawrence Island.*

The covariance estimates in Figure 4.2.4 show some differences between the products. The NASA team and ASI products have larger correlation length scales of about 400 km in radius (the shown box covers an area of 1600 km × 1600 km) where the CCI covariance seems to be more localized. The NASA team product has a stronger connection across the Bering Strait then the ASI product and shows in addition lower covariance along the coasts.

A major question is how reliable these estimates are; the assumption of independent errors of the products might not hold. To investigate the robustness of our approach we use a fourth SIC product (the Bootstrap Algorithm from the NSIDC doi: https://doi.org/10.7265/N59P2ZTG) and repeat the above analysis with each possible pair of three. We investigate the dependency of the CCI+ SIC product error covariance estimate on the choice of the two other products which are used to derive it. Ideally all three covariance estimates in Figure 4.2.5 would be in good agreement. While there are some consistent features (e.g., reduced CCI+ covariance across the Bering strait and mostly increased covariance near the coast), there are also substantial differences (Figure 4.2.5). This is thought to be caused by a failure of the independence assumption between the SIC products. We derived error covariance estimates for all four products and analyzed other locations and seasons and find that this approach appears to be more suited for the NASA Team and ASI algorithms (higher consistency, not shown). There are two likely reasons for this. (1) CCI+ error correlations are typically weaker which makes it more likely that weaknesses in these estimates are overpowering the signal and/or (2) the cross-product error correlations are in such a way that they do have a stronger impact the CCI+ product (which is not a quality characteristic). In the equations, allowing for a positive cross-product error correlation for two of the three involved SIC products results in an underestimation in the spatial error covariance estimates of the two correlated products and an overestimation of the spatial covariance of the third, independent SIC product.

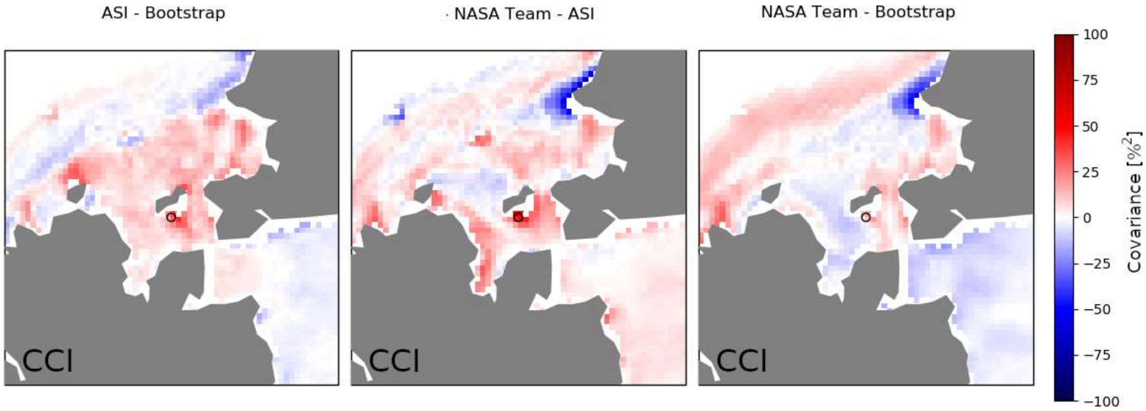


Figure 4.2.5: *Covariance estimates (equations given in Figure 4.2.3) for February of the CCI+ SIC product based on triangulation with three sets of other SIC products (left: SSMI/ASI + Bootstrap, centre: NASA team + SSMI/ASI, right: NASA Team + Bootstrap) relative to the location marked by a black circle, approx. 200 km south of the Bering strait and just north of St. Lawrence Island.*

Synthesis

We developed an SIC error correlation model and attempted to constrain the corresponding parameters based on a statistical analysis of the data. None of the approaches (individual SIC sample correlations, re-analysis of the PVIR circular correlation estimates and a triangulation of error correlations by a combination of several SIC products) lead to a robust estimate of the model parameters. It did, however allow us to improve our understanding of the SIC error characteristics and cross-product correlations. Overall, spatial covariance structures can have significant non-circular components (Figure 4.2.1), correlation length scales are rarely below

100 km and frequently reach several hundred km (Figure 4.2.2, Figure 4.2.5), and there are weak indications for increased covariance pattern in the CCI+ products near land (Figure 4.2.5, centre and right). Our correlation model is capable of incorporating all those findings. The lower bound of correlation length (l_{x0}) should be chosen to be no less than 100 km and the sum of the two additional length scales (l_{xL} and l_{xSLC}) should be at least a few hundred km to cover the whole range of SIC error correlation length scales found (Figures 4.2.1, 4.2.2 and 4.2.5). If simple circular error correlation models are used we would suggest a few hundred kilometres as length scale. Note that neglecting the error correlation when e.g., deriving the SIA uncertainty would result in an implicit decision to set the characteristic error correlation length scale to a value well below the SIC product resolution (zero), which would in general not be in agreement with our results.

Publications

None so far, but we plan to describe related concrete plans in the next version of this report.

Interactions with the ECVs used in this experiment

Interactions between the CMUG and ECV projects for work on this WP in particular happened though an email exchange with CCI+ SI team members (Thomas Lavergne, Stefan Hendricks and Stefan Kern) as well as joining and presenting at meetings, including CCI+ colocation meetings and the CCI+ SI progress monitoring meeting in March 2021.

Consistency between data products

This section will provide a record of any inconsistencies found between ECV products, and will be completed in the next version of this report.

Recommendations to the CCI ECV teams

To be completed in next version of this report.

4.3 Evaluation of model results considering the abstraction level of observational products

Lead partner: MPI-M

Author: Andreas Wernecke

Aim

The aim of this research is to develop and apply a framework that allows one to estimate the ideal abstraction level at which a model evaluation should be carried out. It will address the following scientific question: At which observational abstraction level should we evaluate large-scale model simulations?

Summary of Work and Results

The abstraction level of observational data has many layers. Taking the example of sea ice, ‘observations’ can refer to anything from the sensor measurements themselves to the strength of a process, like the observed ice mass flux through the Fram Strait. Generally speaking, model evaluations can be done at a level close to what is measured (which we call a small abstraction level), at a level close to model variables (a more traditional approach) or at the process level (high abstraction) which can be extracted from both observational products and models.

In the traditional approach, measurements of satellite sensors (here we focus on satellite-based measurements due to the exceptional importance for climate model evaluations) are processed towards physical properties like e.g., the area fraction of the ocean covered by sea ice (Sea Ice Concentration, SIC). Processing includes sensor calibration and georeferencing, geometrical and signal interference corrections, corrections for extraterrestrial radiation and atmospheric influences as well as attribution of portions of the signal to different surface types (including snow, land, etc.) and other aspects inferred with the variable retrieved. The latter includes the impact of snow (e.g., structure, temperature and thickness), ocean state (e.g., wave spectrum, surface films) and land (e.g., surface type, temperature) on the sensor measurements and hence the retrieval of the SIC. In practice it is often necessary to use auxiliary data for those corrections (such as near surface air temperature to identify likely surface melt and corresponding changes to the snow/ice properties) and to interpolate, including gap-filling in the observational products for a systematic assessment of models. These and more factors should ideally be represented by uncertainty estimates provided with the products.

The idea behind using smaller abstraction levels for model evaluation is to use model variables to simulate what a given state of the modeled system would look like at the satellite sensors. These models are also called observation operators. There is a large potential benefit in this approach because many of the before mentioned corrections have their root in the climate system and can therefore be part of the modeled system.

One example is the role of the atmosphere in passive microwave SIC estimates. If satellite sensor measurements are converted to SIC estimates, the impact of the atmosphere must be assumed to be small or corrected by additional data (e.g., climatologies), which will often be uncertain themselves. A coupled GCM simulates the atmosphere as well as the ocean and cryosphere and can hence make estimates of how the microwave signal should look at the satellite, if the climate system as a whole is well represented. On a technical note, atmospheric water vapor/ice clouds play only a minor role for most passive microwave frequencies used for SIC estimates (which is one reason why they are used), but it has some influence at the near 90 GHz channel and indirect influence on e.g., the ocean roughness and surface temperatures. Another example is the ice thickness for which the altimeter measured ice freeboard (distance between ice surface and free ocean) is converted into an ice thickness using, among other things, (rough) estimates of the snow load on top of the ice. Within models the current, local snow thickness can be simulated together with the ice thickness, so that a model freeboard can be easily calculated and compared with the (less erroneous) measured ice freeboard. An off-side of using small abstraction levels is the reduced interpret ability. In the last example, a mismatch between modeled and observed freeboard does not allow for conclusions whether the ice-, or the snow-thickness is wrong which would be a valuable information for model improvements. In addition, known model biases in one variable can spread into all parts of the model evaluation. The electromagnetic (emission/transfer/reflection) models used to translate model variables into sensor level estimates have uncertainties themselves. In addition, climate models do not have a direct representation of many aspects the real climate system. This might include aspects which are important for operation operators (see e.g., the topic of meltponds in Burgard et al. 2020b). In Table 4.3.1 and Table 4.3.3 we attempt to summarize the before-mentioned factors as model uncertainty.

Even though sensor level model evaluation has been successful in other fields (including studying the formation of galaxies [Bower et al. 2010]), it is a newly evolving topic in the field of remote sensing of sea ice [Richter et al. 2018, Burgard et al. 2020a, b]. Radar altimeter return shapes (waveforms) have been modeled for freeboard retrievals [Kurtz et al. 2014], but not for model evaluation.

On the other extreme, deriving estimates of a process strength (e.g., Fram Strait ice flux, Atlantic meridional overturning circulation etc.) can be very effective to reduce observational uncertainties and be highly informative for model development. However, it is often not possible to identify a single process which is a good metric for model quality in general. The potential for process-based evaluations is therefore strongly dependent on the application, so that we restrict ourselves here to high abstraction level measures which are directly related to model variable fields and are widely applicable.

Table 4.3.1: Overview of uncertainties in observations and from observation operators (Model unc.) related to Sea-ice concentrations. All values are rough estimates and vary from satellite/product/study to satellite/product/study as well as regionally and over time.

Abstraction Level	Quantity	Units	Observation unc.	Model unc.
Sensor	Microwaves TB	K	<1%	<=10% (1, 7)
	Thermal infrared TB	K	<1%	Unknown
Area fraction	SIC	%	<10% (2, 3, 8)	0
	Potential open water fraction	%	~10% (5)	Unknown
Hemispheric measures	SIA	km²	<10% (March), <25% (Sept.) (3, 6)	0
	SIE	km²	<5% (March), <10% (Sept.) (3)	<1% (March), <10% (Sept.) (3)

Table 4.3.2: Complementary information to Table 4.3.1

	Notes to Table 4.3.1
1	Errors from GCM simplifications vs. high resolution model (Burgard et al. 2020a) and vs assimilated SIC (Burgard et al. 2020b) typically below 10K TB
2	Ivanova et al. 2015; SIC uncertainty frequently reaches ~40% in the marginal ice zone due to interpolation errors (see reference in note 4)
3	Notz 2014
4	SICCI Phase 2 SIC Product User Guide (SICCI-PUG-P2-17-09)
5	Drüe and Heinemann 2004
6	Most of this uncertainty is caused by biases, trends are more certain
7	Differences in TB due to transfer model up to 15K, Richter et al. 2018
8	CCI+ Sea Ice ECV Sea Ice Concentration PVIR (D4.1)

Table 4.3.3: Overview of uncertainties in observations and from observation operators (Model unc.) related to radar-based Sea-ice thickness estimates. The focus on radars is because of the availability of long, continuous time series which are essential for model evaluations. All values are rough estimates and vary from satellite/product/study to satellite/product/study as well as regionally and over time.

Abstraction Level	Quantity	Units	Observational unc.	Model unc.
Sensor	Echo return	W	<15cm (1,3)	Unknown (2)
Vertical extend on 25kmX25km grid	Ice freeboard	m	<=10cm (3, 4)	0
	Draft	m	<1m (4, 5)	0
	SIT	m	<=1m (3, 4)	0

Table 4.3.4: Complementary information to Table 4.3.3

	Notes to Table 4.3.2
1	Wingham et al. 2006, for individual measurements
2	Kurtz et al. 2014 model Cryosat-2 Waveformes and fit them to measurements but do not asses the uncertainties in the waveform modelling or sensitivity to variables whichare not covered in GCMs. To our knowledge there is no such assessment in the literature.
3	Ricker et al. 2014
4	CCI+ Sea Ice ECV SIT PVIR
5	Reported ice draft RMSE with validation products are largely smaller than 1 m but very low correlation coefficients and poor apparent performance in Fig. 15 in the SIT PVIR make us increase this uncertainty. SIT PVIR (Section 3.2.1) reports a RMSE of 55 cm with submarine measurements and 14 cm with buoy data

As can be seen from Tables 4.3.1 and 4.3.3, there is so far no proven reduction in uncertainties by using observation operators since the added uncertainties from linking the model variables to observed satellite signals are similar to the uncertainties in traditional observational products (SIC and SIT). It should be mentioned that the potential for improvement of observation operators is large. That is because they are just evolving and because each generation of climate models is expected to cover more aspects of the climate system which makes it likely that observation operators can improve in simulating the sensor signals.

The ice freeboard, however, has a smaller observation uncertainty than the ice thickness. This is expected to remain the case when considering relative uncertainties (the freeboard is smaller than the ice thickness). Climate models routinely simulate sea ice thickness, snow cover/thickness and have a water, ice and snow density, so that a climate model consistent ice freeboard can be easily calculated. The absent model uncertainty in Table 4.3.2 for ice freeboard indicates that all variables are available in the model to exactly calculate the ice freeboard, it does not mean that said variables values match real world values. This leads to the main reason why one might want to consider using the ice thickness for model evaluation instead of the ice freeboard (despite its smaller uncertainty). When comparing observed and modelled freeboard it is not clear whether mismatches originate from a poor representation of sea ice in general or for example from errors in the local snow density. While the first would probably be a reason for concern, the latter might be acceptable.

In summary, currently the only reduction in observational uncertainties which is not offset by uncertainties in observation operators, we could find here, is the use of ice freeboard instead of ice thickness. Whether the ice freeboard should be used for model evaluation further depends on the need for interpretability in the model evaluation.

Publications

None so far, but the interest in the results leading to a journal or conference publication will be described in the next version of this report.

Interactions with the ECVs used in this experiment

There have been interactions with the SSS, Snow, SST, SI and LST CCI ECV projects at the quarterly CSWG meetings and the Integration meetings. We have been in regular contact with the CCI SI team, attending progress meetings and through the preparation of the Sea Ice Climate Assessment Report.

Consistency between data products

No inconsistencies found so far.

Recommendations to the CCI ECV teams

To be completed in next version of this report.

4.4 Optimal spatial and temporal scales for model evaluation

Lead partner: MPI-M

Authors: Andreas Wernecke

Aim

The aim of this research is to develop and apply a framework that allows one to estimate the ideal spatial and temporal time horizon at which a model evaluation should be carried out to minimize the impact of observational uncertainty. It will address the following scientific question: At which time and space scale should we evaluate large-scale model simulations?

Summary of Work and Results

So far, our work on this topic addresses optimal spatial and temporal scales separately. This is in line with common practice where typically spatial averaging is done for each time slice, and temporal averaging is either done for each location separately or on globally aggregated quantities. However, the results presented here are nevertheless expected to be informative for simultaneous spatio-temporal averaging. The concept of optimal scales is clearly dependent on the application. If interested in evaluating seasonal process representation in a model, shorter averaging periods are necessary than for decadal evaluations. We will hence focus on the possible gains for model evaluations by averaging as a function of reduced information, which can be assessed by the Degrees of Freedom (DoF) of a dataset. In other words, if most gains (e.g., reduced observational uncertainties) manifest by averaging on short scales, the optimal scale for model evaluation will be shorter as well since less loss of information/DoF has to be tolerated.

Spatial scales

Here we borrow the concept of a Fractions Skill Score (FSS) from evaluations of rain forecast models. The idea is that the exact locations of rain (typically defined by a threshold on the precipitation) in a forecast might be mis-located for local rain events even if the regional forecast is of good quality. In this case the model would be unnecessarily penalised on a grid-cell by grid-cell evaluation even if performing well overall. Sea ice is comparable to rain locations in that it is binary on a very small scale and that we do not expect global models to correctly reproduce the exact locations of each patch of ice. On the other extreme, focusing solely on global measures (like SIA) might unnecessarily discard valid information. The FSS is a straightforward measure of the quality of agreement over all spatial resolutions. It is the normalized mean squared error between two fields which is calculated for increasing levels of spatial signal smoothing. This smoothing is realized by a mean-filter with increasing box size from the original resolution (where the mean filter has no effect) to box sizes larger than the domain, where each location is assigned the domain mean value. We adjust the calculation for the influence of land, so that land covered cells have no influence on the calculation of the nearby mean-filtered SIC. In the following we will first test the behaviour of the FSS on a toy example representing an ice floe attached at different locations to an ice front. After that we apply the FSS on MPI-ESM model simulations in a perfect model approach to investigate how inter-model differences reduce with effective resolution and finally use CCI+ SIC observations as an example of model evaluation with the FSS.

Sea ice fractions skill score – proof of concept

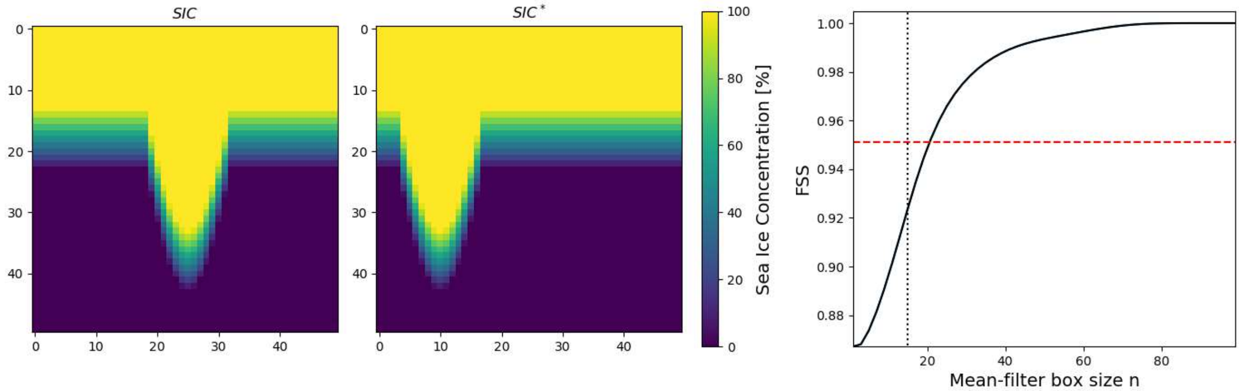


***Figure 4.4.1:** Fractions Skill Score (FSS, right) for a test case (left) in which an ice tongue/floe at the edge of a pack ice field has been moved by 15 cells to the left (compare left and centre). The black dotted line shows this 15-cell offset in the FSS panel and the red dashed line indicates a proposed threshold where the FSS increased by 63% from the n=1 to the maximal FSS.*

Figure 4.4.1. shows a test case for applying the FSS on sea ice. Here the main feature in the SIC field is offset by 15 grid cells and the FSS quickly increases around this spatial scale. We also highlight the FSS for which the difference between the FSS and max(FSS) has fallen to 1/e from the initial value (red dashed line in Figure 4.4.1, right panel). The size of the filter box which crosses this threshold (here 21 grid cells) is used as estimate of where most of the gain in FSS is achieved. To better understand the meaning of this measure, called ñ, we address test cases similar to Figure 4.4.1 but with a range of different ice tongue offsets and derive ñ for each of them.

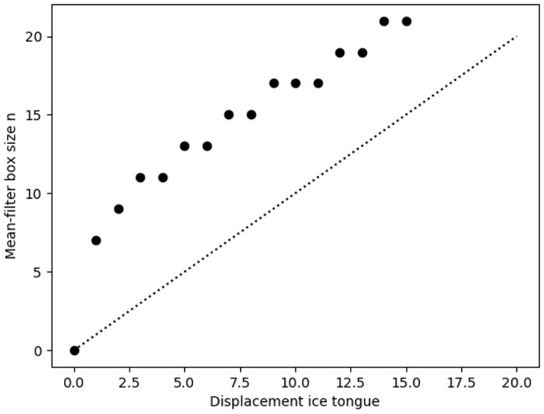


***Figure 4.4.2:** Mean-filter box size for which FSS increased at least 63% from size=1 to max(FSS) (ñ, y-axis) versus displacement of ice tongue/floe in number of boxes (x-axis).*

As can be seen in Figure 4.4.2, the value of ñ increases approximately linearly with the displacement of the ice tongue (for displacements>0) with a slope of ~1 and a positive y-axis offset of about 5. This y-axis offset corresponds to about half of the ice tongue width, which we also find for other widths (not shown).

In summary, the spatial scale at which the FSS increases quickly depends bi-linearly on the mean offset between patches of ice and their typical sizes. If those two factors are small, a comparison of the two underlying fields can be done with higher spatial resolution since local features have more constrained size and location.

Before we apply the FSS to compare model and satellite observations we will establish a baseline purely from climate model data. This is done to see on which scales MPI-ESM grand ensemble (GE) SICs begin to harmonize as an internal model characteristic before we use the FSS to investigate the agreement between satellite observations and model simulations. Thereby we will get a better understanding of how well model results can be expected to agree with observations on a large number of scales.

Sea ice fractions skill score – perfect model test

We use the MPI-ESM GE member #1 as reference and derive the FSS. In the following we selected northern hemisphere February and September data (near the yearly sea ice maximum and minimum) of 2005 as examples. We brought all datasets to a 50 km X 50 km north polar stereographic grid.

Figure 4.4.3 shows the FSS between 99 MPI-ESM GE simulations and the selected reference. It can be seen that most of the gain in FSS is achieved in the first ~1500 km, allowing the following conclusions. The typical size of features in the SIC data which are displaced is below 1500 km. In addition, the distance those patterns are displaced is mostly below 1500 km as well. It is worth mentioning that the difference in total SIA (color coded in Figure 4.4.3) shows no clear relation to high or low FSS (maybe with the exception of two outliers in September), supporting the notion that the analyzed information (distance and size of displaced pattern) is indeed independent information from the total SIA and as such provides additional value for model evaluations. While the shape in FSS gain is largely in agreement between February and September, the absolute FSS is smaller in September. To reach a specific FSS, larger spatial averaging intervals have to be used in September than in February. This is could well be related to the fact that the coast of the Arctic Ocean is becoming a fixed boundary for the sea ice in winter creating higher spatial agreements. Figure 4.4.3 also includes the Degrees of Freedom (DoF), approximated by the number of initial ocean grid cells divided by the grid cells used in the moving window of the mean-filter. While the DoFs drop sharply with box size, it should be noted that, e.g., for a box side length of 1000 km, there are still approximately 100 DoF. Compared with just one DoF for hemisphere wide measures and about 10 if the Arctic is divided into its marginal seas, this is still a relatively large number. This figure further provides the level of agreement which can be expected for different model simulations at a given resolution of interest.

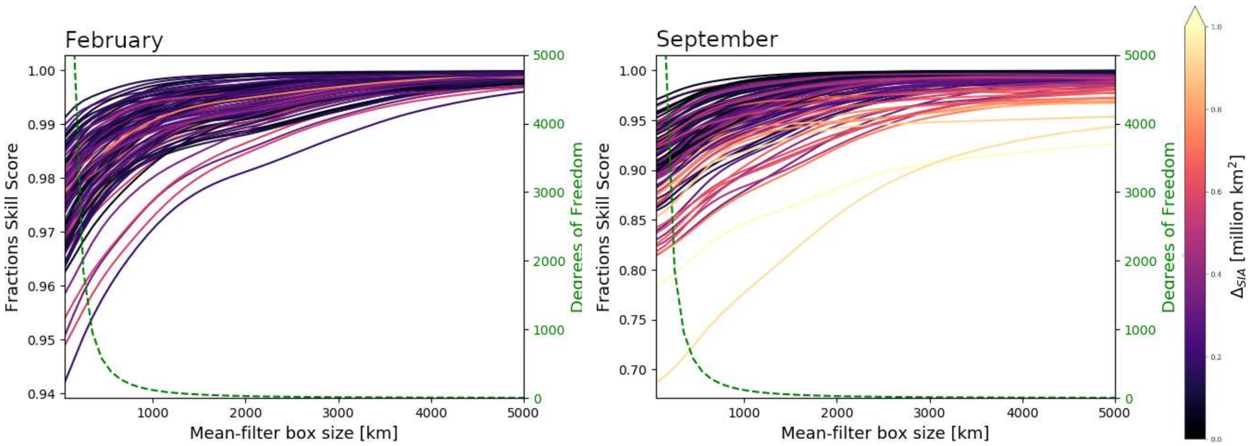


***Figure 4.4.3:** Fractions skill score of model simulations relative to a selected reference model simulation (#1) for northern hemisphere SIC in February (left) and September (right). Also included are the DoF with effectively degraded spatial resolution (green dashed line) and the difference in total sea ice area (color coded).*

Finally we calculate ñ for each simulation and provide the resulting frequency distribution in Figure 4.4.4. The distribution of ñ does not change notably between February and September.

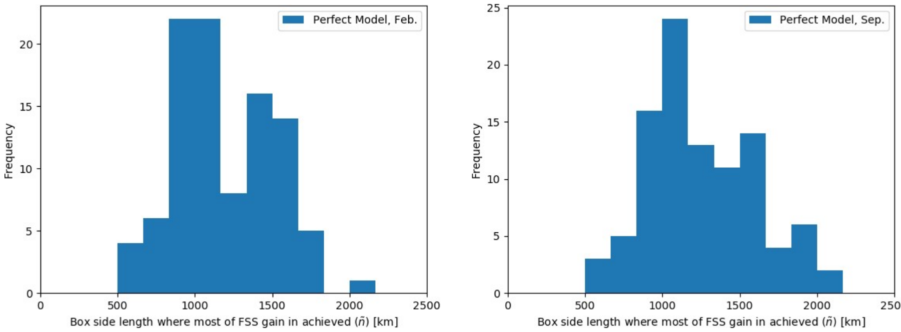


***Figure 4.4.4**: Frequency distribution of ñ for February (left) and September (right) with one simulation as reference.*

Sea ice fractions skill score – CCI+ SIC with model simulation

We now repeat the above exercise with monthly mean CCI+ SIC data as reference. In this case (Figure 4.4.5 and Figure 4.4.6) the February and September FSS show more differences. The February FSS has slightly lower values for small box sizes (I.e., high spacial resolution) compared to the previous inter-model results (Figure 4.4.3 left). This could be related to the fact that the native model grid has a smaller resolution or that physical processes near the model resolution are not represented as well. In both cases the regriding to 50 km X 50 km effectively oversamples the data. Comparing the satellite data with a real resolution close to 50 km with smoother model data reduces the FSS at the lower end of Figure 4.4.5 (left). This reduction in

FSS at the lower end of mean-filter box size also explains the smaller values for ñ. It summarizes the fact that, due to the mentioned effective oversampling, model evaluations benefit more by spatial averaging on relative short scales.

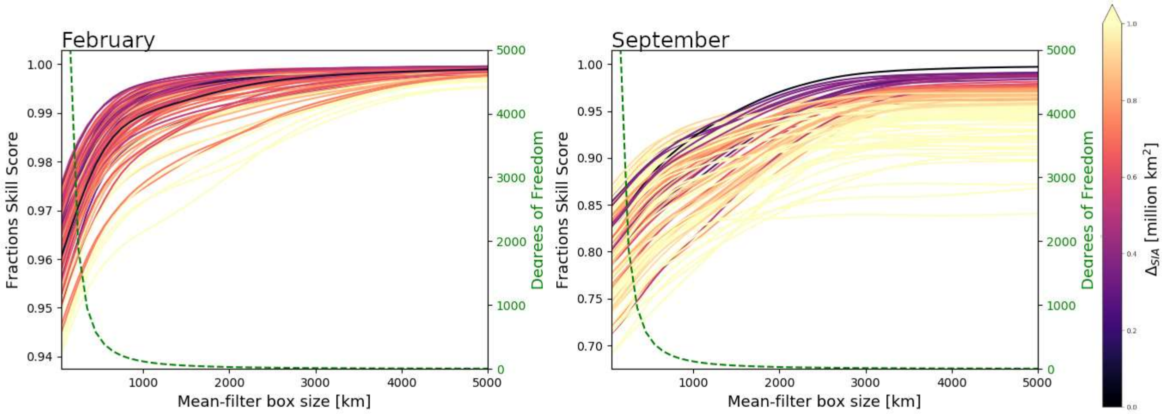


Figure 4.4.5: *Fractions skill score of model simulations relative to northern hemisphere CCI+ SIC observations in February (left) and September (right). Also included are the DoF with effectively degraded spatial resolution (green dashed line) and the difference in total sea ice area (color coded).*

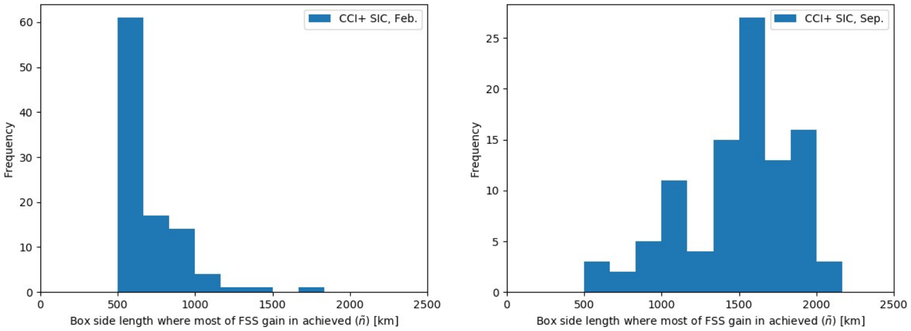


Figure 4.4.6: *Frequency distribution of ñ for February (left) and September (right) with CCI+ SIC observations as reference*

The September estimates have much larger differences in total SIA (the modelled SIA is smaller than the observed, not shown) and require more spatial averaging to reach elevated FSSs. This points at a more substantial mismatch between model and observations. Figure 4.4.7 illustrates this mismatch where the difference in SIA (approximately 0.8 million km²) is high but not worse than some inter simulation differences (Figure 4.4.3). The spatial distribution is clearly worse with the simulation showing large areas with intermediate SIC between 40% and 80% and the observations show a smaller extent with higher concentrations. The FSS hence correctly identifies pattern mismatches of concern, identifies simulations which are less prone to these (i.e., the simulations with higher FSS in Figure 4.4.5, right) and suggests that larger spatial scales are needed for meaningful model evaluations around the yearly sea ice minimum than around the sea ice maximum.

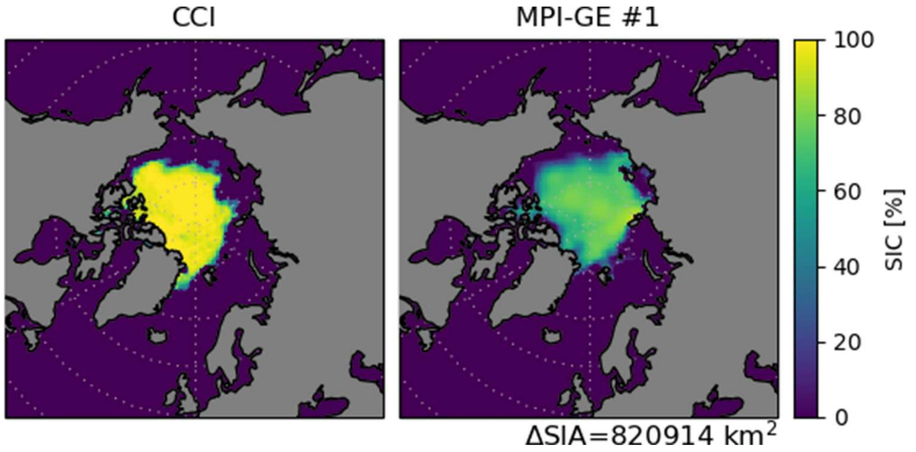


Figure 4.4.7: *Sea-ice concentration for September 2005 from CCI+ satellite observations (left) and MPI-ESM GE run #1 (right).*

Temporal scales

For the evaluation of optimal temporal scales, we follow a similar approach as for spatial scales. We start with the full data resolution (daily) and monitor the development of quality metrics while degrading the resolution. Instead of a spatial mean-filter on time slices we use averages of temporal bins in the SIA anomaly time series. To derive the anomaly, we use the daily SIA, subtract the yearly cycle and linear trend. The quality metrics are the correlation coefficients and standard deviations of time series. The correlation coefficient requires a common signal in the time series to be meaningful. Here, the signal is the variability in the SIA which is not the same for any two long term climate model simulations. Instead, we use observational datasets which share the same real SIA variability as the signal and differ by observational uncertainty. The products used are the NASA-Team, Bootstrap and NASA-Merged algorithms (all processed by NASA) as well as the OSI-SAF algorithm. The NASA-Merged product is a combination of NASA-Team and Bootstrap algorithms and the OSI-SAF data are used here instead of the CCI product due to the longer period covered (but they are closely related for the period when they overlap).

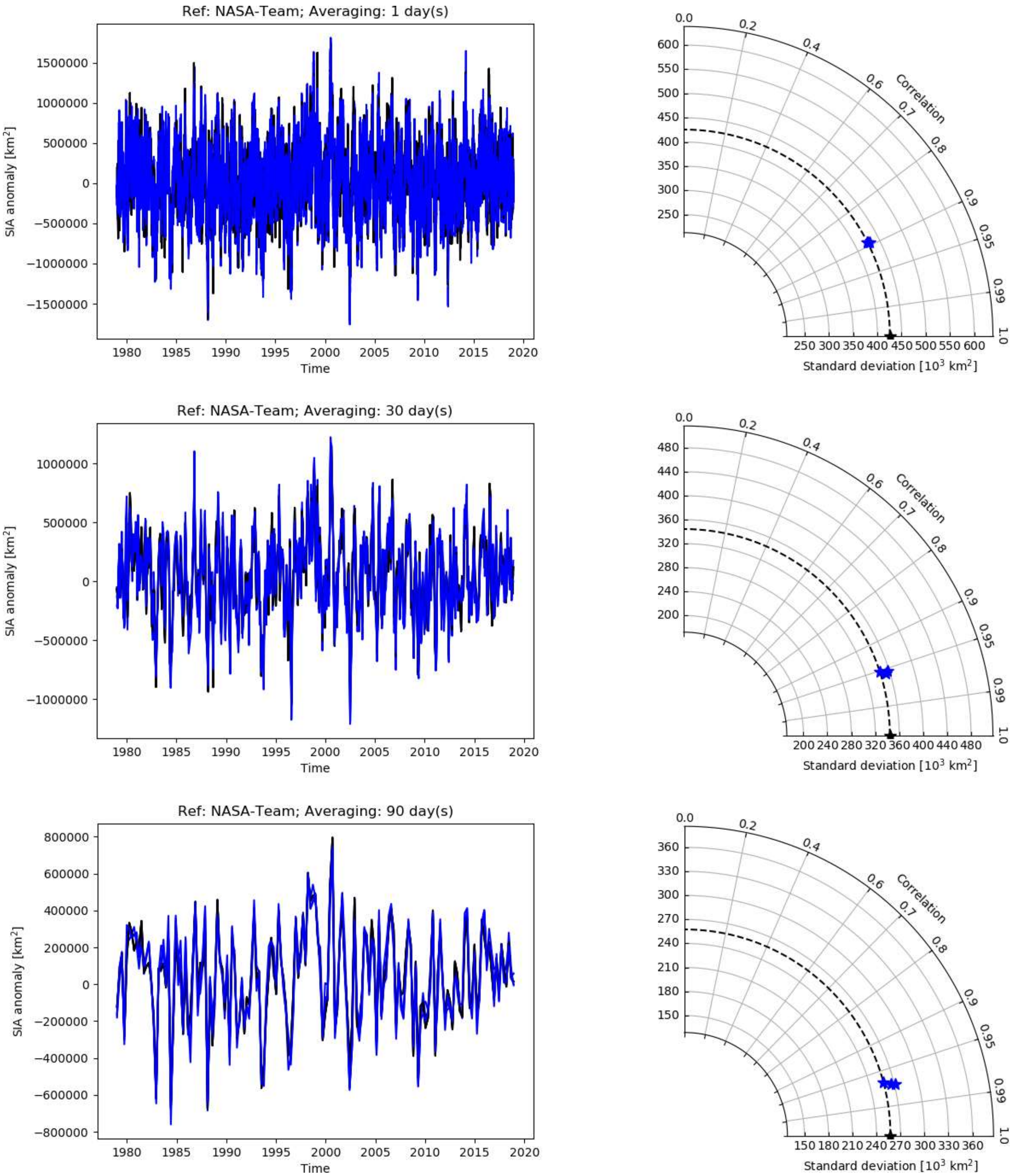


Figure 4.4.8: *Time series (left) and Taylor-diagrams (right) between different observational SIA data products (shown are detrended anomalies). The NASA-Team product has been selected as reference (black lines and star) but other choices show the same picture. The temporal resolution has been degraded from the original daily (where available, top) to 30-day averages (middle) or 90-day averages (bottom).*

As can be seen in Figure 4.4.8, the fluctuations in the SIA time series reduces with degraded resolution (note also the different axis ranges for the standard deviation on the Taylor diagrams) which can be attributed to reduced observational uncertainties as well as a loss of resolved real SIA variability. The correlation between the SIA products in Figure 4.4.8 increases which shows that the reduction in uncertainty is larger than the loss in resolved variability. The optimal temporal scale for model evaluation would ideally be such that observational uncertainty is averaged out but the signal (I.e., real SIA variability) is preserved. While such an ideal scale will not exist, we have a look at the relationship between standard deviation and correlation for a range of temporal scales in below.

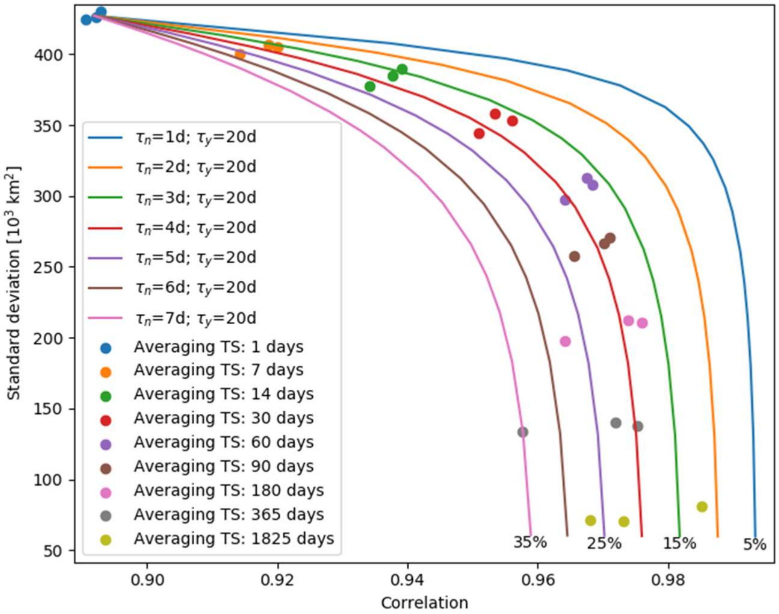


Figure 4.4.9: *Standard deviation vs. correlation coefficient towards NASA-Team SIA time series of the three remaining products for several averaging periods (dots). For comparison, analytical results (lines) for which the uncertainty and signal are considered autoregressive models of order one with different temporal correlation length scales for the observational uncertainty (tau_n) and constant temporal correlation length scale for the real SIA variability (tau_y), here selected to be 20 days. The percentage values correspond to the fraction of uncertainty- to signal length scales. See text for more information.*

When increasing the averaging window, the correlation is strongly increasing at first with moderate losses in anomaly fluctuations up until 30 to 60 days (Figure 4.4.9). After this larger window sizes result in reduced anomaly fluctuations with no further gains in correlation. From this we conclude that temporal averages can be useful (depending of course on the application) to reduce observational uncertainties up until about two months. After this no additional advantage is expected.

For a better understanding of the results, we derived analytical solutions for a statistical representation of these time series. For this we assume the SIA anomaly and observational uncertainties to be independent and represent both by separate zero-mean autoregressive models with order one (AR(1)). These have a white noise term and a memory term so that we can prescribe the amplitude of fluctuations and the characteristic autocorrelation length scale. Observational products are represented by the sum of the uncertainty AR(1) and the signal AR(1). Since any two products have independent, identical uncertainty models but share the signal model, we can use conditional likelihoods to derive the analytical correlation between them. For AR(1) models we can also find the standard derivation and correlations of temporal averages. For Figure 4.4.9 we force the statistical models to match the standard deviation and correlation of the observations for the daily data and set a range of temporal correlation length scales manually. The behavior for increasing averaging windows is in first order defined by the ratio of the two autocorrelation length scales (not shown).

The results from the observations largely follow the theoretical results with uncertainty autocorrelation of 3 to 5 days for signal autocorrelation of 20 days. This corresponds to uncertainty autocorrelation length of 15% to 25% of the signal autocorrelation length. A similar relation is also found for other theoretical signal length scales (not shown). This suggests that the observational uncertainty autocorrelation scale is likely to be about one fifth of the real SIA anomaly autocorrelation scale.

To summarize this WP, we analyzed the impact of spatial and temporal SIC degradations on the quality of model and observational agreement. For MPI-ESM GE simulations we find that most of the benefits from spatial averaging are realized below 1500 km and we can provide users with an estimate of expected model quality spread for averaging on any specific scale. Evaluating model data (long term climate simulations without data assimilation or initialization in SIC observational period) with CCI SIC observations show that substantially larger averaging intervals are needed around the yearly sea ice minimum than around the sea ice maximum. In other words, the medium to high resolution spatial agreement in SIC is considerable worse in September than in February.

For temporal averaging we find that the time series fluctuations (consisting of error fluctuations and real variability) reduce moderately for averaging periods of up to about two months. At the same time the correlation between SIA products increases, while for longer averaging periods the correlation is not further benefitting but the fluctuations keep reducing. This indicates that at first the errors reduce, after which more and more of the real signal is lost. With the help of theoretical considerations, we estimate that the temporal autocorrelation of observational uncertainties is expected to be around one fifth of the length scale of the SIA variability. If the typical period of a process of interest is known, it should therefore be sufficient to base any analysis on temporal averages with a resolution of a fifth to a quarter of the period of interest to minimize the influence of observational uncertainties.

Publications

None so far, but the interest in the results leading to a journal or conference publication will be described in the next version of this report.

Interactions with the ECVs used in this experiment

In the first 12 months of this phase of CMUG work there have been interactions with the SSSal, Snow, SST, SI and LST CCI ECV projects at the quarterly CSWG meetings and the Integration meetings. Contact outside that has been only to check on the continuation of the SI and SST projects, and to learn about the beta data that LST announced was available in late 2019. Interactions with the SIMIP project are planned for the future.

Consistency between data products

This section will provide a record of any inconsistencies found between ECV products, and will be completed in the next version of this report.

Recommendations to the CCI ECV teams

To be completed in next version of this report.

4.5 Evaluation of model results considering internal variability

Lead partner: MPI-M

Authors: Dirk Olonscheck, Dirk Notz

Aim

The aim of this research is to develop and apply a framework that allows one to consider the impact of internal variability into a model-evaluation processing chain. It will address the following scientific question: How can we take internal variability into account when evaluating large-scale model simulations?

Summary of Work and Results

The work done in the first year of this CMUG research period (October 2018 to September 2019) was on the methodology that will be used on the new CCI+ datasets when they are available. The method allows one to easily take model-specific internal variability into account when evaluating simulations from global climate models. This lays the methodological basis for taking internal climate variability into account when evaluating climate-model simulations with the forthcoming CCI+ ECVs. The background research on which the CMUG work is based was published in Olonscheck and Notz, 2017, and an evaluation using CMIP5 simulations from that paper is shown in Fig. 4.5.1.

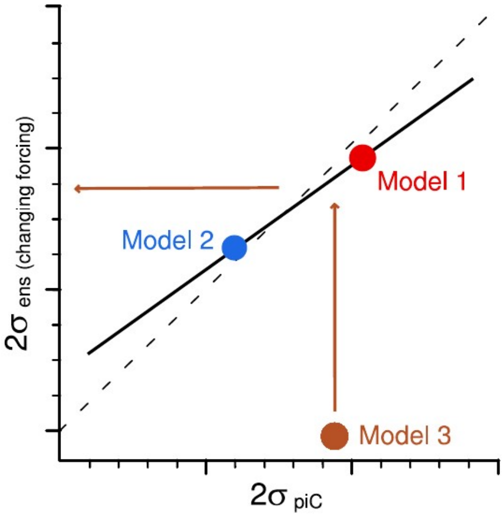


Figure 4.5.1: *Schematic view of the method for estimating internal variability for different forcing scenarios.*

The basic version of the method regresses the estimate of internal variability derived from the preindustrial control simulation of a model (x axis) on the ensemble standard deviation of models with ensemble simulations such as models 1 and 2 (y axis). The unity line as a reference is indicated by the dashed black line. For the extended version, a constructed ensemble standard deviation can be derived for models with a single simulation (model 3) using the regression line through models 1 and 2. The extended version requires a consistent response of the models with ensemble simulations. A summary of the scientific outcomes of the research are:

1. Development of a new method that allows us to consistently estimate internal climate variability and its change over time for all models within a multimodel ensemble such as CMIP5 by regressing each model's estimate of internal variability from the preindustrial control simulation on the variability derived from a model's ensemble simulations.

2. We find a highly variable model-specific internal variability of sea-ice volume and sea-ice area.

3. The method allows for the evaluation of climate-model simulations by uniformly taking model-specific internal variability for all models into account.

Publications

None so far, but the interest in the results leading to a journal or conference publication will be described in the next version of this report.

Interactions with the ECVs used in this experiment

In the first 12 months of this phase of CMUG work there have been interactions with the SSSal, Snow, SST, SI and LST CCI ECV projects at the quarterly CSWG meetings and the Integration meetings. Contact outside that has been only to check on the continuation of the SI and SST projects, and to learn about the beta data that LST announced was available in late 2019. Interactions with the SIMIP project are planned for the future.

Consistency between data products

This section will provide a record of any inconsistencies found between ECV products, and will be completed in the next version of this report.

Recommendations to the CCI ECV teams

To be completed in next version of this report.

4.6 Evaluation of model results considering a combination of sources of uncertainties

Lead partner: MPI-M

Authors: Dirk Olonscheck, Dirk Notz

Aim

The aim of this research is to develop and apply a framework that allows one to include both observational uncertainty and uncertainty arising from internal variability into a model-evaluation processing chain. It will address the following scientific question: How can we take observational uncertainty and internal variability into account when evaluating large-scale model simulations?

Summary of Work and Results

The work done in the first year of this CMUG research period (October 2018 to September 2019) was on the methodology that will be used on the new CCI+ datasets when they are available. The introduced plausibility variable (below) allows one to take both model-specific internal variability and observational uncertainty into account for evaluating climate-model simulations. We did so to evaluate the CMIP5 climate-model simulations as shown in Fig. 4.6.1. This comprehensive evaluation approach will be applied to comparing climate-model simulations with the CCI+ ECVs. The background research on which the CMUG work is based was published in Olonscheck and Notz (2017).

We introduce a plausibility variable as a measure of model fidelity, which takes both the model-specific internal variability (sigma_mod) and the observational or reanalysis uncertainty (delta_ref) into account:

$$\phi = \frac{\overline{\text{mod}} - \overline{\text{ref}}}{\sqrt{\sigma_{\text{mod}}^2 + \delta_{\text{ref}}^2}}$$

This approach to evaluate climate-model simulations considers both internal variability and observational uncertainty and thus links to Task 4.2.

The results allow for a distinction between model deviations that are plausible due to internal variability and reference-data uncertainty and those that cannot be explained by these sources of uncertainty, pointing to model biases.

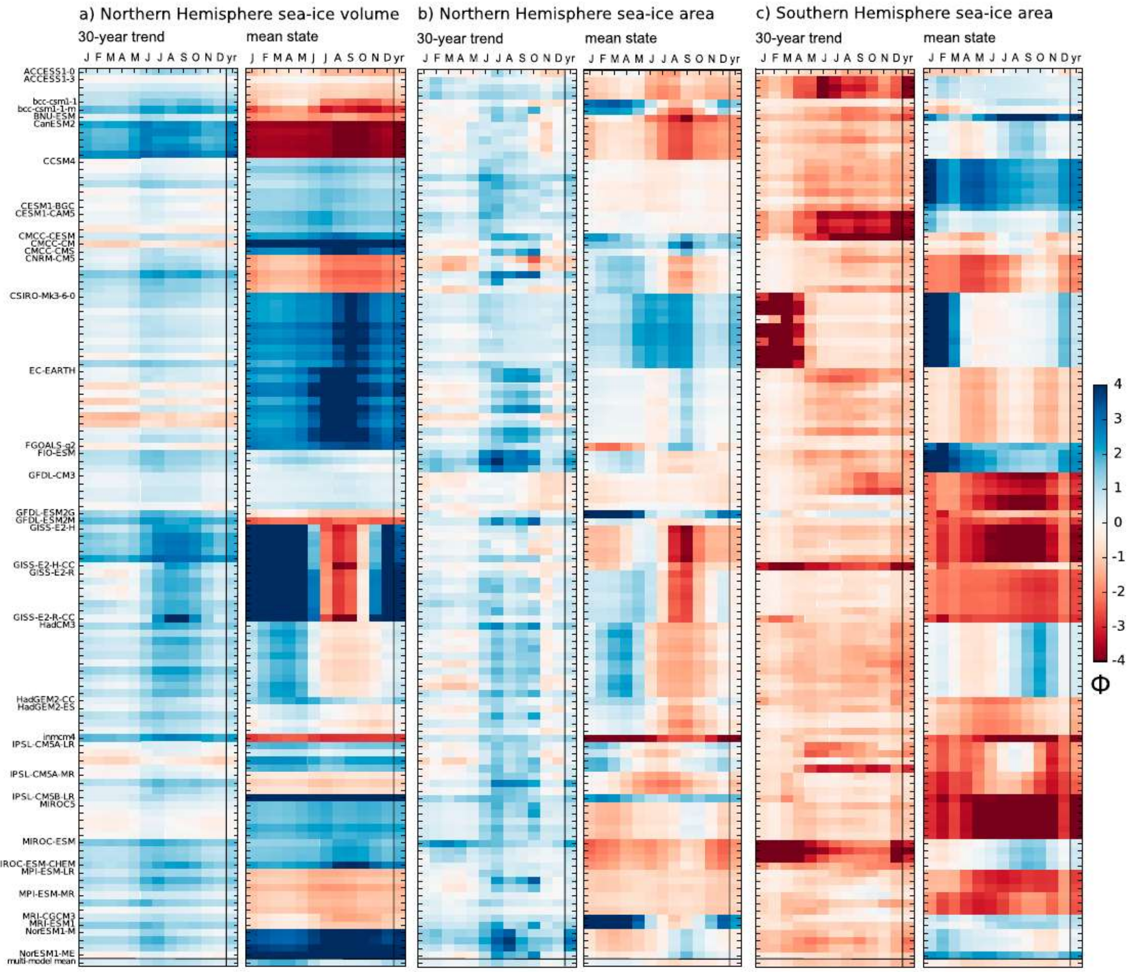


__Figure 4.6.1__: Portrait plot of the plausibility of CMIP5 sea-ice simulations for the 30-yr trend and the mean state of (a) Northern Hemisphere sea-ice volume, (b) Northern Hemisphere sea-ice area, and (c) Southern Hemisphere sea-ice area based on the distance between each extended historical CMIP5 model simulation and reference data (PIOMAS for Northern Hemisphere sea-ice volume and satellite sea ice data from a CCI precursor dataset, Meier 2013, for sea-ice area). Deviations are shown in units of "phi", which combines delta_ref and sigma_mod; a model's negative (red) and positive (blue) deviation with respect to reference data are indicated. Note that each model name is attached to the first ensemble simulation only.

Publications

None so far, but the interest in the results leading to a journal or conference publication will be described in the next version of this report.

Interactions with the ECVs used in this experiment

In the first 12 months of this phase of CMUG work there have been interactions with the SSSal, Snow, SST, SI and LST CCI ECV projects at the quarterly CSWG meetings and the Integration meetings. Contact outside that has been only to check on the continuation of the SI and SST projects, and to learn about the beta data that LST announced was available in late 2019. Interactions with the SIMIP project are planned for the future.

Consistency between data products

This section will provide a record of any inconsistencies found between ECV products, and will be completed in the next version of this report.

Recommendations to the CCI ECV teams

To be completed in next version of this report.

4.7 Skill assessment of the DCPP decadal predictions

Lead partner: BSC

Authors: Roberto Bilbao, Jaume Ruíz de Morales, Froila Palmeiro, Pablo Ortega and Louis-Philippe Caron.

Aim

The aim of this WP is to produce an extensive model skill assessment of the decadal hindcasts done within DCPP (Decadal Climate Prediction Project, Boer *et al.* 2016; thus contributing to CMIP6 initiative) with the longest CCI products available as an independent source for validation, thus testing at the same time the consistency of CCI data with the reference datasets used for their initialization. It will address the following scientific questions:

1. Which are the regions/variables with more skill for decadal prediction across climate models?
2. Can CCI/CCI+ data help to identify if these are robust across datasets?
3. Does skill arise for different variables over the same region?
4. Can this help to identify the processes behind the skill?

Summary of Work and Results

Preliminary skill assessment of the EC-Earth decadal prediction system based on ocean reanalyses and objective analyses:

The skill of decadal prediction systems is typically assessed by performing large sets of retrospective predictions (or reforecasts) that are later contrasted against the observed past variability. For the EC-Earth decadal predictions used in this WP, the reforecast period used goes from 1960 to 2020, with 10-member ensembles of predictions initialised every 1st of November. Because the longest satellite observations are only available, in the best case, since 1979, leaving 20 start dates out of the skill assessment, we started the analysis (version 1 of D4.1) by performing a first assessment against longer ocean datasets, namely the ocean reanalyses and objective analyses: ORAS5, EN4 and HadISST, to account for the observational uncertainty.

The BSC completed the decadal reforecasts by September 2019, and contributed with them to the Decadal Climate Prediction Project (DCPP) component A of the World Climate Research Programme (WCRP). The decadal predictions are performed with a resolution of T255L91 in the atmosphere and 1° and 75 vertical levels in the ocean. The initialization technique used is full-field initialisation. Atmospheric initial conditions were generated using ERA-40 and ERA-Interim. Land initial conditions were taken from ERA-40 prior to 1979. From then onwards ERA-Land was corrected with GPCP observations and used as land initialisation. Ocean and sea-ice initial conditions were produced using a NEMO-only simulation forced by DFS5.2 atmospheric fields and nudged towards ORAS4.

The preliminary analysis of skill in the EC-Earth hindcasts was carried out for monthly-mean global-mean sea surface temperature (SST) (Figure 4.7.1a) and several climate variability indices derived from SST: the Atlantic Multidecadal Variability (AMV) index (Figure 4.7.1c) calculated with the Trenberth and Shea (2006) definition, and El Niño Southern Oscillation (ENSO) index (Figure 4.7.1e) defined as the SST average over the Nino3.4 box (5S-5N and 170-120W). Beforehand, anomalies had been computed from the raw predictions with respect to the period 1970-2005 using a lead-time dependent climatology. To quantify the deterministic skill the anomaly correlation coefficient (ACC) was used (Figure 4.7.1).

The ACC of global-mean SST in EC-Earth hindcasts showed high skill for the 5 first forecast years (Figure 4.7.1b). A large part of the skill of the decadal predictions is associated with the global warming trend, as later demonstrated in Bilbao et al. (2021), by computing the skill in a set of non-initialized CMIP6 historical simulations (DECK+ScenarioMIP) in which only the externally forced changes are represented. Interestingly, Figure 4.7.1d also shows that the model has high skill in reproducing the AMV, which is a mode of internal variability that is not predicted in the historical experiments. The same happens for ENSO, which the DCPP simulations are capable of skilfully predict during the first forecast year. We also note that all of these results are consistent independently of the observational product that is used to compute the skill. A more detailed analysis of the skill in this decadal prediction system with EC-Earth can be found in Bilbao et al (2021).

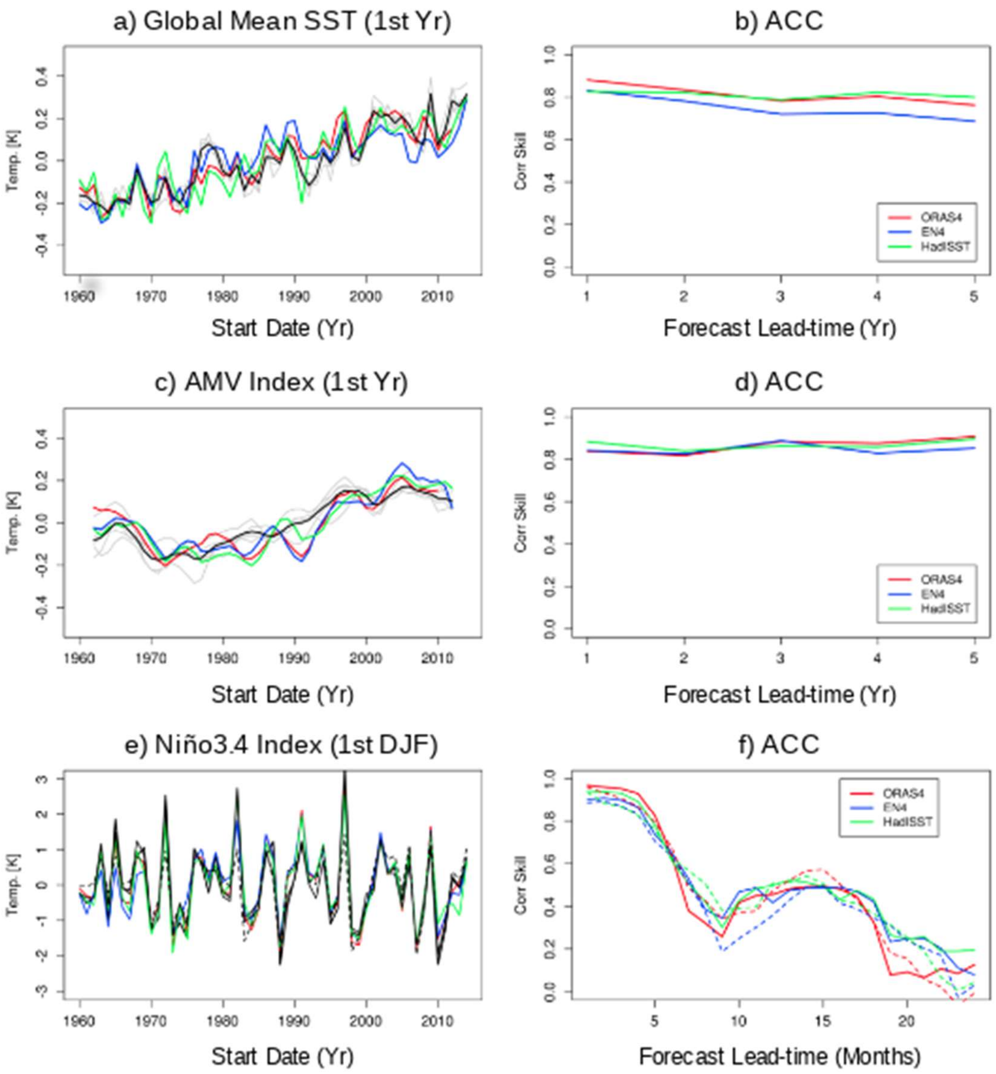


Figure 4.7.1: *a) Annual-mean global mean SST for forecast lead time year 1 for the ensemble mean decadal predictions (black), individual members (grey) and three observational products: ORAS4 (red), EN4 (blue) and HadISST (green)). b) Anomaly correlation coefficient (ACC) of the ensemble mean hindcasts and observations. c) As figure a) but for the AMV (Trenberth and Shea, 2006 definition). d) ACC of the AMV. e) As figure a) for the Niño3.4 index for the first DJF. e) ACC of the Niño3.4 index.*

Decadal Skill assessment against a selection of long and physically related CCI products:

Jaume Ruiz de Morales, a master student from the University of Barcelona, joined the BSC in February 2021 and is conducting a new analysis of the retrospective decadal predictions, focused on a more recent period in which they can be evaluated against satellite products, under the supervision of Roberto Bilbao (who performed the decadal predictions and did the preliminary analysis), Froila Palmeiro and Pablo Ortega. The three main ECVs selected for this analsysis are SST, Sea Level and Cloud Cover, which offer several advantages with respect to the data of other ECVs: 1) they are available for more than 20 years, which is key for computing robust skill metrics, 2) they have more than one observational product available, which allows us to investigate the consistency across them and the sensitivity of the results to the product employed, 3) they relate to variables for which decadal predictability is expected, and 4) they are, at least to some extent, physically related with each other (e.g., SST can influence cloud cover through evaporation and covary with the sea level height in areas in which its thermosteric component is dominant).

The list of products employed and their main characteristics are summarised in the table below:

Variable (units)	**Product**	**Time Period**	**Original resolution**
Cloud cover (%)	EUMETSAT ESA AVHRR-PM v3.0	01/1982-05/2019 01/1982-12/2016	0.25º 0.5º
Sea Surface Height Anomaly (in m)	C3S CMEMS	01/1993-10/2019 01/1993-02/2020	0.25ºx0.25º 0.25ºx0.25º
Sea Surface Temperature (in °C)	ESA L4 HadISSTv1.1 ERSST	01/1982-12/2020 01/1870-12/2020 01/1854-12/2020	0.05º 1ºx1º 2ºx2º

Because not all products cover the same period, for each variable only the period of overlap has been considered when the skill was evaluated. In addition, all data (both for the model and observations) have been regridded to a regular 1°x1° grid for practical reasons. The predictions have been evaluated against a dataset resulting from the average of the different datasets identified for that variable. This was done to constrain the common signals, which are more likely to be true, and average out some of the observational uncertainties, which are expected to be different across products. A more detailed assessment of the consistency across the different products is included in the section "Consistency across data products"

Skill assessment of Sea Surface Temperature: We have begun by evaluating the skill to predict the SST evolution at different forecast ranges (Figure 4.7.2), in terms of the ACC. In the first forecast year, large and significant ACC values are obtained in most regions of the world, with the main exceptions of some Polar areas (in which the sea ice might not have been properly initialised), and a small region in the Central North Atlantic. A similar pattern was obtained in Bilbao et al. (2021) for a longer reforecast period (1960-2018) and against a different observational dataset, which confirms that skill is consistent in time and insensitive to the validation dataset. Bilbao et al. (2021) also showed that most of this skill arises from the externally forced signals, with only a few regions like the Tropical Pacific and the North Atlantic Subpolar Gyre exhibiting skill from internal variability processes. At longer forecast ranges (years 1 to 5 and 6 to 10, respectively) the ACC values are increased, although this is just a consequence of evaluating the forecasts on 5-year means (thus smoothing some of the interannual variability noise). We can see, however, that the Central North Atlantic remains as a region of very poor skill, a problem that has been linked in Bilbao et al. (2021) to an initialization shock in Labrador Sea deep convection. Other regions like the Eastern Tropical Pacific and the Southern Ocean also show very negative ACC values, the first probably reflecting a problem in the model representation of the frequency of ENSO, and the second probably due to a very strong local warm bias in the EC-Earth version used to perform the forecast, which grows with forecast time. Despite these regions of poor skill, most areas of the global ocean show high level of predictability even 6-10 years after initialization, an encouraging result that speaks about the utility of the decadal predictions.

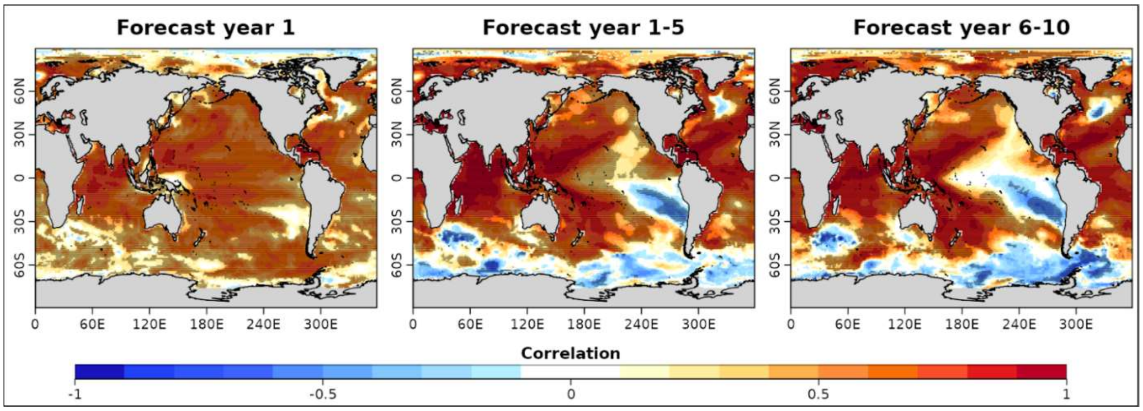


Figure 4.7.2: *Anomaly correlation coefficient (ACC) for the annual sea surface temperature in the CMIP6 decadal prediction system with EC-Earth for forecast years 1 (left panel), 1–5 (middle panel) and 6–10 (right panel). The ACC is computed between the model ensemble mean of the ensemble mean of 3 observational products: ESA L4, HADISSTv1.1 and ERSST. Correlation values that are significant at a 95% confidence level are indicated with stippling. The skill is assessed for the period 1982-2020.*

Skill assessment of Sea Level Height: The related variable that is directly simulated by the model is the dynamic sea level (DSL) as defined in Griffies and Greatbatch (2012), which reflects the sea level fluctuations related to ocean dynamics and is computed as the sea level anomaly with respect to the ocean geoid (which can vary in time, e.g., if the ocean experiences a thermal expansion). By construction this variable has a global mean of zero in every time step. To assess the model ability to predict it, we have therefore converted the observations of the Sea Surface Height Anomaly (in which the anomaly is computed against a temporal mean instead of a spatial mean) to this same quantity by removing the global mean. Figure 4.7.3 shows very high levels of skill to predict the DSL in the first forecast year both in the Pacific and Indian basins. Areas of significant skill are very scarce in the Atlantic, and mostly concentrate in the subtropics and the Labrador Sea. At subsequent forecast years (1-5) most of the skill is lost, and the major areas of significant skill are the Indian Ocean and the subtropical Pacific. Most of the skill is lost in those regions at the longest forecast years (6-10), although the Labrador Sea show some promising levels of predictability. It is interesting to note that, apart from the first forecast year, the areas of significant skill are different than for SST, which reflects that the processes yielding the predictive skill are different in each case. The same analysis has been repeated in Figure 4.7.4, but focused on the absolute sea surface height anomaly (i.e. the variable directly measured by satellites). It is important to note that in this case this involved adding the global mean sea level to the dynamical sea level within the model. We also note that since this version of EC-Earth does not resolve the contributions from continental ice sheets and glaciers, sea level changes due to meltwater fluxes are not simulated, and only the thermosteric component is resolved (and can be potentially predicted). The ACC figure of the absolute sea level changes shows much higher and persistent skill values than for the DSL, particularly large in the Tropical regions. This most probably reflects that the global thermosteric component (excluded in Figure 4.7.3), which mostly consists of an increasing linear trend and is mostly anthropogenic, is highly predictable in the model.

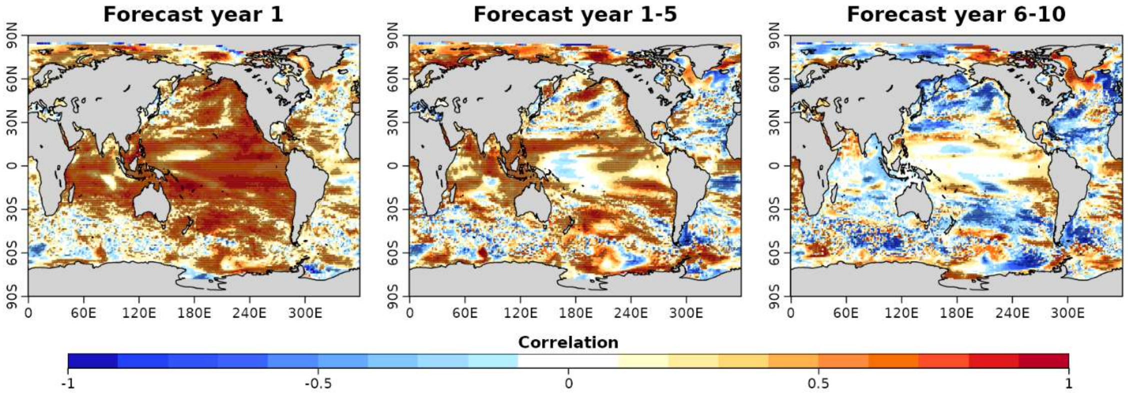


Figure 4.7.3: The same as in Fig. 4.7.2 but for the dynamic sea level.

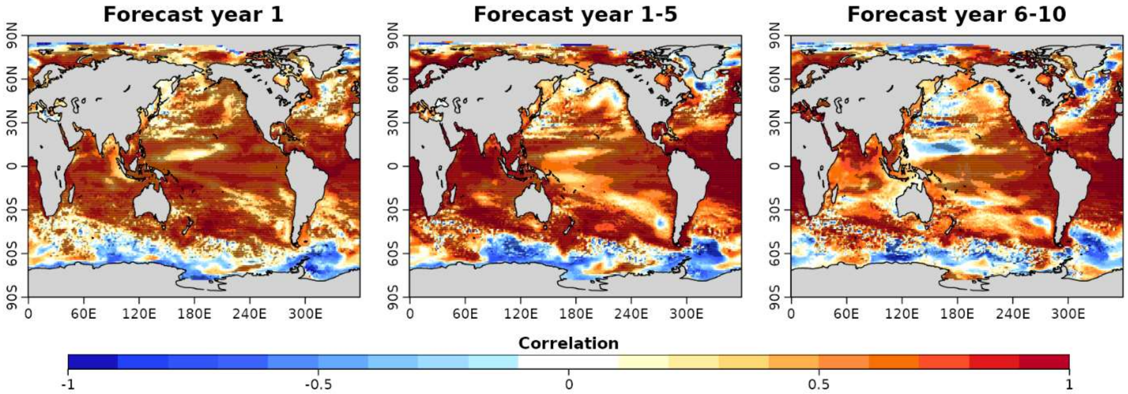


Figure 4.7.4: The same as in Fig. 4.7.3 but for the absolute sea surface height anomaly.

Skill assessment of Cloud Cover: This analysis has not been finalised and will be included in the next version of the deliverable

Publications

A scientific article documenting the production of the decadal prediction system with EC-Earth, evaluating its skill against different reanalysis products, and detailing some encountered issues has been published in 2021 in Earth System Dynamics: *Bilbao, R., Wild, S., Ortega, P., Acosta-Navarro, J., Arsouze, T., Bretonnière, P.-A., Caron, L.-P., Castrillo, M., Cruz-García, R., Cvijanovic, I., Doblas-Reyes, F. J., Donat, M., Dutra, E., Echevarría, P., Ho, A.-C., Loosveldt-Tomas, S., Moreno-Chamarro, E., Pérez-Zanon, N., Ramos, A., Ruprich-Robert, Y., Sicardi, V., Tourigny, E., and Vegas-Regidor, J.: Assessment of a full-field initialized decadal climate prediction system with the CMIP6 version of EC-Earth, Earth Syst. Dynam., 12, 173–196, https://doi.org/10.5194/esd-12-173-2021, 2021.*

Interactions with the ECVs used in this experiment

The specific satellite products used in this analysis for the cloud cover and sea level variables were been recommended, respectively, by Martin Stengel (DWD) and Jean-François Legeais (CLS), when we asked them for advice before the skill assessment was started. The ESA product used for SST was presented and recommended by Christopher Merchant in the CSWG meeting devoted to SST, SSS and Sea Ice.

Consistency between data products

To assess the temporal consistency between the different observational products considered, we have computed, for their overlap period, the correlations between each pair of products. Figure 4.8.4 shows the minimum value of such correlations at the grid point level for the three variables considered. Overall, correlations are encouragingly large in most regions of the world, which is indicative of strong consistency across products (and suggestive of small observational uncertainties in terms of interannual variations). The weakest correlations are seen for SSTs in the Southern Ocean, something expected given that this is a region where in-situ observations (which are considered in two of the datasets compared) are very scarce, and therefore larger observational uncertainties remain. Interestingly, we can also see in Figure 4.7.5 b,d,f that when the long-term trends are excluded (a feature in which the datasets can more easily agree), consistency across products remains very high, which suggests that uncertainties in the interannual variability are also small.

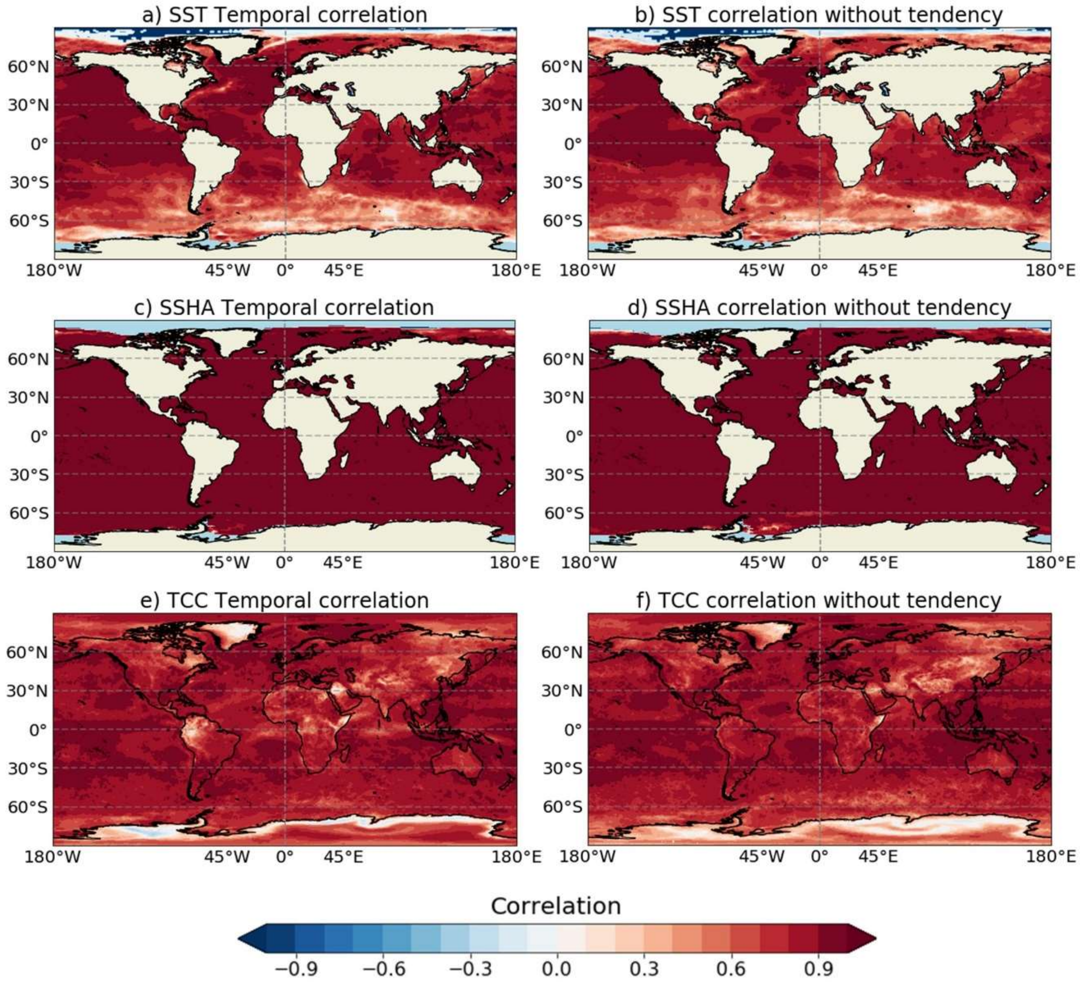

a) SST Temporal correlation
b) SST correlation without tendency
c) SSHA Temporal correlation
d) SSHA correlation without tendency
e) TCC Temporal correlation
f) TCC correlation without tendency
60°N
30°N
0°
30°S
60°S
180°W
45°W
0°
45°E
180°E
Correlation
−0.9
−0.6
−0.3
0.0
0.3
0.6
0.9

Figure 4.7.5: *a-b) Minimum point-wise correlation between the three different SST observational datasets considered, respectively for the raw and linearly detrended annual means. c-d, e-f) The same as in a-b but for the datasets of sea surface height anomaly, and the total cloud cover.*

Recommendations to the CCI ECV teams

To be completed in next version of this report.

4.8 Use LST products to develop and test simple models relating the LST versus air temperature (near surface) difference to vegetation moisture stress

Lead partner: Met Office

Authors: Rob King, Deborah Hemming

Aim

The aims of this research are to: 1) use the differences between LST and Temperature (near surface) to assess spatial and temporal variations in vegetation moisture stress across biomes. SM will also be used to examine the vegetation moisture stress. The biomes will be characterised by AGBiomass and LC. 2) Understand relationships between LST and Temperature in the context of vegetation carbon exchanges across biomes and regions. 3) Assess the potential for using LST versus Temperature relationships as a large-scale monitor of vegetation moisture stress. It will address the following scientific questions:

1. Can LST versus Temperature relationships be used to monitor large-scale vegetation moisture stress across different biomes and regions?

2. What quality information can be learned from the ancillary ECVs used in this study?

Key features

1. Understand how LST to (near surface) air temperature differences vary across biomes and to different biomes.

2. Understand how vegetation moisture stress effects LST to (near surface) air temperature differences to give indicators of moisture stress for particular biomes.

Summary of Work and Results

We have carried out some preliminary investigations and familiarisation with the soil moisture CCI data that is currently available and have started looking at the beta CCI LST data that has just become available.

Publications

None so far, but the interest in the results leading to a journal or conference publication will be described in the next version of this report.

Interactions with the ECVs used in this experiment

In the first 12 months of this phase of CMUG work there have been interactions with the LST, AGB, SM and LC CCI ECV projects at the quarterly CSWG meetings and the Integration meetings. Contact outside that has been with LST who have already produced a beta dataset, and AGB (by email) to discuss data specifications and access, and the wider question of coordinated work on using the AGB data in a land surface climate model. Interactions with SM and LC have been to learn about the continuation datasets they will be producing in CCI+. Interactions with the LUMIP and Decadal Climate Prediction projects are planned for 2020.

Consistency between data products

This section will provide a record of any inconsistencies found between ECV products, and will be completed in the next version of this report.

Recommendations to the CCI ECV teams

To be completed in next version of this report.

4.9 Use CCI+ products and simple models developed in WP4.8 to evaluate performance of modelled LST versus air temperature, using multiple up-to-date land surface and Earth System models

Lead partner: Met Office

Authors: Rob King, Deborah Hemming

Aim

The aim of this research is to evaluate how well the observed relationships between LST and Temperature across different vegetation types and moisture regimes are captured by the JULES land surface model, UKESM1 and other CMIP5 and 6 (where available) Earth System Models. It will address the following scientific question:

1. Can models capture the LST versus Temperature (near surface) relationships observed with satellite products across different vegetation types and moisture regimes?

Key features

1. Identify biome specific relationships between LST and near-surface air temperatures in LST CCI data

2. Evaluate the models (listed above) in their LST and air temperature, to understand how they capture the relationship seen in the CCI data. This evaluation will cover different biomes to capture both differing vegetation types (land cover) and (soil) moisture regimes.

Summary of Work and Results

We have some insights from a preliminary investigation about the behaviour of JULES in particular biomes when skin temperatures (LST) are compared with the driving air temperatures.

Publications

None so far, but a paper on the evaluation of modelled seasonality in vegetation is planned.

Interactions with the ECVs used in this experiment

In the first 12 months of this phase of CMUG work there have been interactions with the LST, AGB, SM and LC CCI ECV projects at the quarterly CSWG meetings and the Integration meetings. Contact outside that has been with LST who have already produced a beta dataset, and AGB (by email) to discuss data specifications and access, and the wider question of coordinated work on using the AGB data in a land surface climate model. Interactions with SM and LC have been to learn about the continuation datasets they will be producing in CCI+. Interactions with the LS3MIP, C4MIP, LUMIP and Decadal Climate Prediction projects are planned for 2020.

Consistency between data products

This section will provide a record of any inconsistencies found between ECV products, and will be completed in the next version of this report.

Recommendations to the CCI ECV teams

To be completed in next version of this report.

4.10 Comparison of CCI products for studying vegetation variations with other satellite products and land surface models

Lead partner: Met Office

Authors: Rob King, Deborah Hemming

Aim

The aims of this research are to: 1) Compare the seasonal timing and magnitude of vegetation-relevant CCI products with other satellite products (including MODIS) and vegetation variables from existing historic model runs (of JULES, UKESM1, CMIP5/6). 2) Identify significant differences in the timing, location and vegetation types between CCI products and other satellite and model data. 3) Suggest key areas for model development to improve vegetation seasonality. 4) Contribute results to a multi-model evaluation conducted in the CRESCENDO project. It will address the following scientific question:

1. Can the large-scale CCI ECV satellite products be used to improve representation of sensitivities and thresholds between vegetation productivity (and other carbon cycle processes) and climate in land surface/Earth System Models?

Key features

- Evaluate modelled vegetation phenology (seasonal timing and magnitude) for JULES UKESM1 and CMIP5/6 historic runs using CCI (and other e.g., MODIS) vegetation products.

- Contribute to multi-model ensemble evaluation for CRESCENDO project.

Summary of Work and Results

Preliminary evaluation of vegetation phenology peak of season modelled with CRESCENDO project models has been conducted using Leaf Area Index monthly products from MODIS and Copernicus Global Land Surface (GLS). Results show significant differences (up to 5) in the magnitude, and variations (of 1-3 months) in the timing of peak LAI between models. Models showed generally later peaks in LAI than the MODIS and GLS satellite products, which were consistent with each other. Other vegetation variables, including Biomass CCI, will be used to assess the magnitude and timing of peak productivity. Initial contact has been made with the Biomass CCI project lead, and a review of the Biomass CCI reports - Product Validation Plan and Uncertainty Budget, was submitted as part of other CMUG work.

Publications

None so far, but a paper on the evaluation of modelled seasonality in vegetation is planned.

Interactions with the ECVs used in this experiment

In the first 12 months of this phase of CMUG work there have been interactions with the LST, AGB, SM and LC CCI ECV projects at the quarterly CSWG meetings and the Integration meetings. Contact outside that has been with LST who have already produced a beta dataset, and AGB (by email) to discuss data specifications and access, and the wider question of coordinated work on using the AGB data in a land surface climate model. Interactions with SM and LC have been to learn about the continuation datasets they will be producing in CCI+. Interactions with the LS3MIP, C4MIP, LUMIP and Decadal Climate Prediction projects are planned for 2020.

Consistency between data products

This section will provide a record of any inconsistencies found between ECV products, and will be completed in the next version of this report.

Recommendations to the CCI ECV teams

To be completed in next version of this report.

4.11 Assess the land-surface interaction related biases in AMIP simulations with CCI and other products

Lead partner: IPSL

Author: Frederique Cheruy

Aim

The aim of this research is to identify biases in the surface state and surface fluxes in AMIP simulations and understanding the origin of these biases in present day simulations (temperature, albedo, fluxes). It will address the following scientific question: What is the potential for exploring multiple satellite derived products to try to relate existing and identified biases (surface state and surface fluxes) to missing or incorrectly represented processes, thus offering solutions for model improvement by revisiting the process representation?

Summary of Work and Results

Role of the input soil textures of the LSM in the regional SM distribution of IPSL-CM

The soil-moisture atmosphere couplings have been assessed for the IPSL-CM in AMIP configuration. An evaluation of the snow cover has been done with alternative products since the snow product was not yet available. The new versions of atmospheric and soil physics of the IPSL model implemented for CMIP6 leads to an image of the interactions between soil moisture and atmosphere that is more consistent with observations. This is particularly true in "hot-spot" regions of strong land-atmosphere coupling and for the driest soils where evaporation and precipitation distributions are closer to those of the observations for the driest soil moisture quartile. Spurious multi-modality in the regional distribution of the superficial soil moisture has been documented over some regions, and is probably related to contrasted field capacities and wilting points as a function of soil texture in our land surface model. This multi-modality is not present in the CCI product, which needs to be investigated by comparing SSM spatio-temporal variability in the three ESA CCI SM products: active (in % saturation), passive (in m3/m3) and combined (which imposes the dynamic range of the GLDAS-Noah SSM product, making this product unfit for bias and RMSD analyses, *cf.* Dorigo et al, 2017). The effect of input soil texture on the distributions of SSM in the IPSL model can also be explored owing to a set of idealized simulations with uniform soil texture over land, recently performed for the Soil Parameter MIP international project (Tafasca et al., 2020)

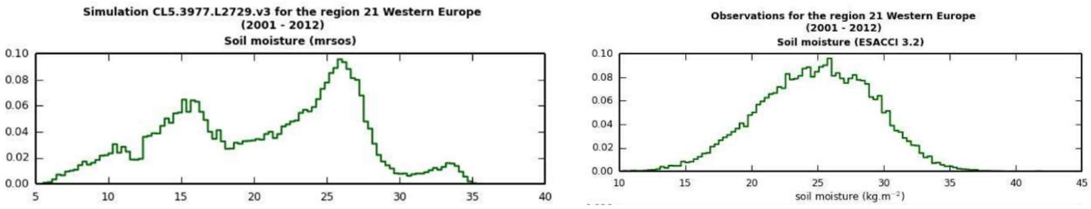


Figure 4.11.1: *Histograms of the soil moisture for the Western Europe as simulated by coupled atmosphere land-surface components of IPSL-CM (left) and retrieved from the CCI product (right).*

While in the real world a large variation in texture is present in each model grid box, the soil hydrology module of IPSL-CM works with the predominant soil texture in each grid box. Only the predominant soil texture is selected for the description of the water fluxes by the soil hydrology module; the tri-modal structure of the histogram reveals the signature of the three different textures present in the Western Europe region according to the USGS-produced soil property maps.

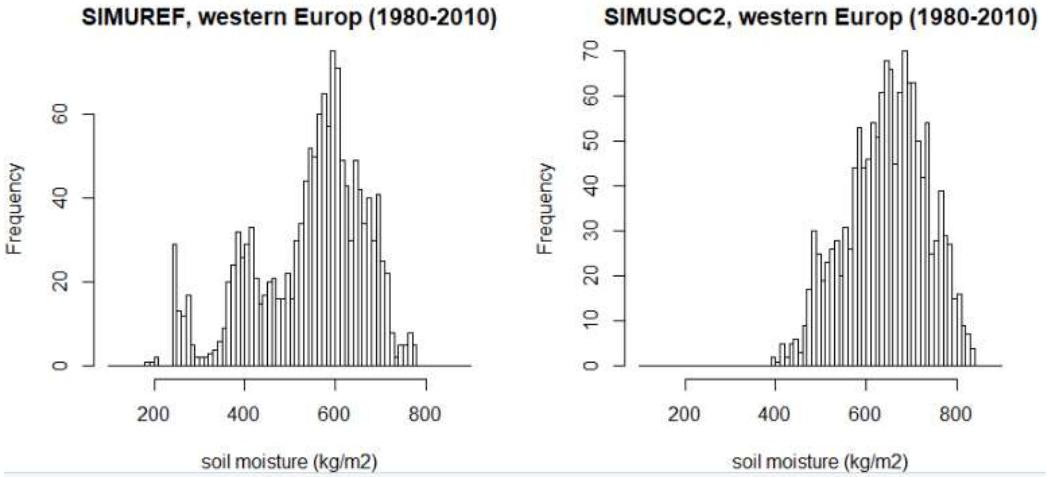


Figure 4.11.2:

The multimodality has been explored and it has been shown that with continuous soil transfer functions (which no longer define the soil properties with a value per texture, but with a real function depending on different parameters, in our case texture and organic matter), no longer has the multi-modality is no longer present in the PDF of the SSM

Realism of the Heat waves and possible biases in the climate models

The realisms of heatwaves as simulated in the AMIP-CMIP6 database has been investigated thanks to various sets of observation-based datasets (table 4.11.1). The multimodel analysis done on a regional basis (Figure 4.11.3) shows that several models exhibit a drier bias during the heatwave with respect to no heatwave days in summer. Both CCI and SMOS show this dry bias. Similarly, heat wave days are associated with an evaporation deficit and a too dry surface layer (relative humidity). These different biases are consistent with an overestimation of the maximum temperature of heat waves at the regional scale (Figure 4.11.4)

Dataset	Variable	Period
Daily data (HadGHCND)	T2max/T2min	1979-2014
Berkeley Earth Surface Temperatures (BEST)	T2max/T2min	1979-2014
ESA CCI SSM COMBINED Product (fv04.5)	SM	1979-2014
GLEAM, v3.5a	E/SM	1980-2014
Soil Moisture and Ocean Salinity (SMOS)	SM	2010-2014
ERA5	T2min/T2max/RH	1979-2014

Table 1: dataset used for the analysis of the realism of the eatwaves in the AMIP- database

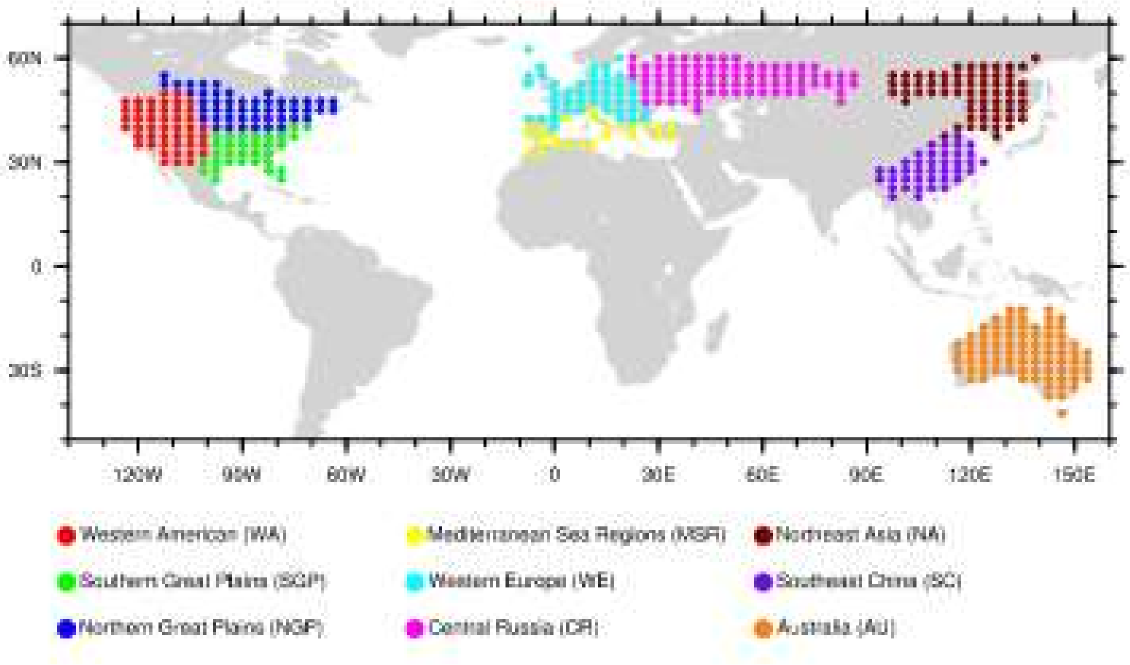


Figure 4.11.3: *Regions used for the heatwave analysis. The regions are based on the Köppen climate classification but only regions were the daily air temperature is correctly sampled are considered.*

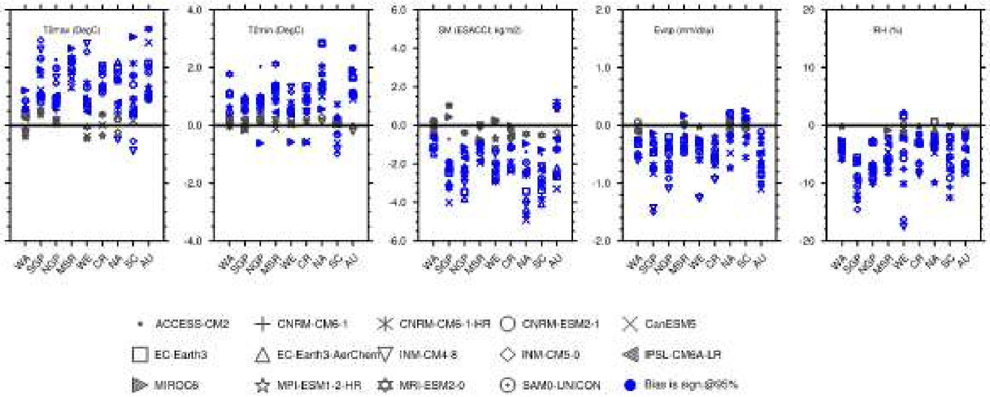


Figure 4.11.4: *Bias difference between HW days and not HW days in summer ([Model_HW-OBS_HW]-[Model_NotHW-OBS_NotHW]) during 1980-2014 for T2max/T2min (□C), with respect to HadGHCND observation), SM (kg/m2, with repect to SM of ESA_CCI), E (mm/day, with respect tto GEAM) and RH (%, with respect to RH of ERA5) for the selected regions. The blue means that the differences are significant at 95% level of confidence, for the 35 years long period. Each marker corresponds to one selected model.*

Publications

None so far, but the interest in the results leading to a journal or conference publication will be described in the next version of this report.

Interactions with the ECVs used in this experiment

In the first 12 months of this phase of CMUG work there have been interactions with the LST and Snow CCI ECV projects at the quarterly CSWG meetings and the Integration meetings. Contact outside that has been with LST who have already produced a beta dataset. Interactions with the LS3MIP, SPMIP, AMIP, HighResMIP and SnowMIP projects are planned for 2021.

Consistency between data products

The comparison of the CCI product to the SMOS product (Figure 4.11.5) shows a moist bias of the order of 0.1 m3/m3 which corresponds to a bias of the order of 10 kg/m2 in the upper 10 cm of soil.

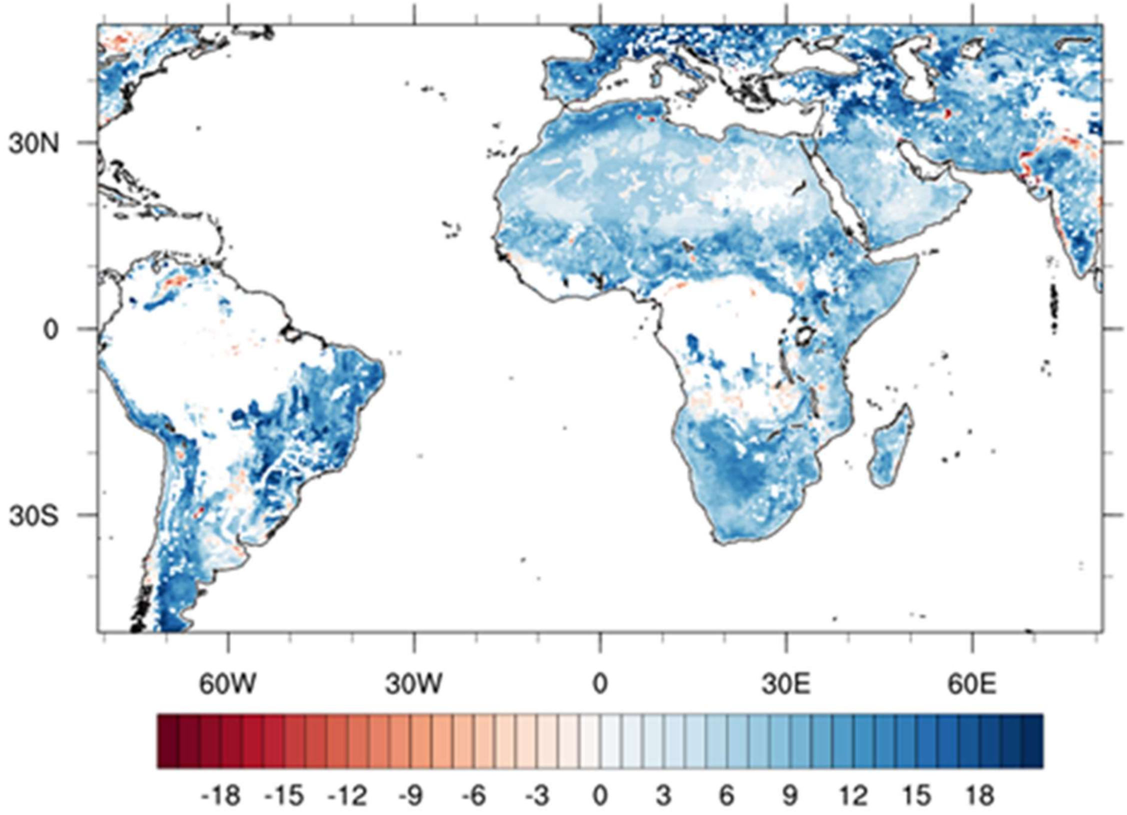


Figure 4.11.5*: Volumetric Soil Moisture difference (SM_CCI – SM_SMOS, unit: 100*m3/m3)*

Recommendations to the CCI ECV teams

To be completed in next version of this report.

5. References

Bilbao, R., Wild, S., Ortega, P., Acosta-Navarro, J., Arsouze, T., Bretonnière, P.-A., Caron, L.-P., Castrillo, M., Cruz-García, R., Cvijanovic, I., Doblas-Reyes, F. J., Donat, M., Dutra, E., Echevarría, P., Ho, A.-C., Loosveldt-Tomas, S., Moreno-Chamarro, E., Pérez-Zanon, N., Ramos, A., Ruprich-Robert, Y., Sicardi, V., Tourigny, E., and Vegas-Regidor, J.: Assessment of a full-field initialized decadal climate prediction system with the CMIP6 version of EC-Earth, Earth Syst. Dynam., 12, 173–196, https://doi.org/10.5194/esd-12-173-2021, 2021.

Boer J., *et al*., 2016: The Decadal Climate Prediction Project (DCPP) contribution to CMIP6, Geosci. Model Dev., 9, 3751–377

Bower, Richard G., Michael Goldstein, and Ian Vernon. "Galaxy formation: a Bayesian uncertainty analysis." Bayesian analysis 5.4 (2010): 619-669.

Burgard, Clara, et al. "The Arctic Ocean Observation Operator for 6.9 GHz (ARC3O)–Part 1: How to obtain sea ice brightness temperatures at 6.9 GHz from climate model output." The Cryosphere 14.7 (2020a): 2369-2386.

Burgard, Clara, et al. "The Arctic Ocean Observation Operator for 6.9 GHz (ARC3O)–Part 2: Development and evaluation." The Cryosphere 14.7 (2020b): 2387-2407.

Dorigo, W., Wolfgang Wagner, Clement Albergel, Franziska Albrecht, Gianpaolo Balsamo, Luca Brocca, Daniel Chung, Martin Ertl, Matthias Forkel, Alexander Gruber, Eva Haas, Paul D. Hamer, Martin Hirschi, Jaakko Ikonen, Richard de Jeu, Richard Kidd, William Lahoz, Yi Y. Liu, Diego Miralles, Thomas Mistelbauer, Nadine Nicolai-Shaw, Robert Parinussa, Chiara Pratola, Christoph Reimer, Robin van der Schalie, Sonia I. Seneviratne, Tuomo Smolander, Pascal Lecomte (2017). ESA CCI Soil Moisture for improved Earth system understanding: State-of-the art and future directions, Remote Sensing of Environment, Volume 203, Pages 185-215, https://doi.org/10.1016/j.rse.2017.07.001.

Drüe, Clemens, and Günther Heinemann. "High-resolution maps of the sea-ice concentration from MODIS satellite data." Geophysical research letters 31.20 (2004).

Griffies, S. M. and Greatbatch, R. J. (2012) Physical processes that impact the evolution of global mean sea level in ocean climate models, Ocean Model., 51, 37–72, doi:10.1016/j.ocemod.2012.04.003

Ivanova, Natalia, et al. "Inter-comparison and evaluation of sea ice algorithms: towards further identification of challenges and optimal approach using passive microwave observations." The Cryosphere 9.5 (2015): 1797-1817.

Kurtz, Nathan T., N. Galin, and M. Studinger. "An improved CryoSat-2 sea ice freeboard retrieval algorithm through the use of waveform fitting." The Cryosphere 8.4 (2014): 1217-1237.

Meier, W., F. Fetterer, M. Savoie, S. Mallory, R. Duerr, and J. Stroeve, 2013: NOAA/NSIDC Climate Data Record of Passive Microwave Sea Ice Concentration, version 2. National Snow and Ice Data Centre, Boulder, CO, accessed 9 December 2015, https://doi.org/10.7265/N55M63M1.

Notz, D. "Sea-ice extent and its trend provide limited metrics of model performance." The Cryosphere 8.1 (2014): 229-243.

Olonscheck, D. & Notz, D. (2017), Consistently Estimating Internal Climate Variability from Climate Model Simulations. Journal of Climate 30, 9555–9573, doi:10.1175/JCLI-D-16-0428.1.

Richter, Friedrich, et al. "Arctic sea ice signatures: L-band brightness temperature sensitivity comparison using two radiation transfer models." The Cryosphere 12.3 (2018): 921-933.

Ricker, Robert, et al. "Sensitivity of CryoSat-2 Arctic sea-ice freeboard and thickness on radar-waveform interpretation." The Cryosphere 8.4 (2014): 1607-1622.

Tafasca, Salma (2020). Evaluation de l'impact des propriétés du sol sur l'hydrologie simulée dans le modèle ORCHIDEE, Sorbonne Université. PhD Thesis.

Tafasca, S., Ducharne, A., and Valentin, C. (2020). Weak sensitivity of the terrestrial water budget to global soil texture maps in the ORCHIDEE land surface model, Hydrol. Earth Syst. Sci. Discuss., DOI: 10.5194/hess-24-3753-2020

Trenberth, K. E. & Shea, D. J. (2006) Atlantic hurricanes and natural variability in 2005. Geophysical Research Letters 33, L12704, doi:10.1029/2006GL026894

Wilks, D. (2019). Statistical Methods in the Atmospheric Sciences, 4th Edition. Elsevier, pp 840. ISBN: 9780128158234

Wingham, D. J., et al. "CryoSat: A mission to determine the fluctuations in Earth's land and marine ice fields." Advances in Space Research 37.4 (2006): 841-871.