H2020 - Research and Innovation Action

APPLICATE

APPLICATE

Advanced Prediction in Polar regions and beyond: Modelling, observing system design and LInkages associated with a Changing Arctic climaTE

Grant Agreement No: 727862

Deliverable No. 1.1

Model Assessment Plan

Version 2, November 2018

Submission of Deliverable

Work Package	WP1 Weather and climate model evaluation		
Deliverable No	1.1		
Deliverable title	Model Assessment Plan		
Version	2		
Status	Final		
Dissemination level	PU - Public		
Lead Beneficiary	1 - AWI		
Contributors	☐ 1 – AWI	☐ 2 – BSC	☐ 3 - ECMWF
	☐ 4 – UiB	☐ 5 – UNI Research	☐ 6 – MET Norway
	☐ 7 – Met Office	☐ 8 – UCL	☐ 9 - UREAD
	☐ 10 – SU	☐ 11 – CNRS-GAME	☐ 12 - CERFACS
	☐ 13 – AP	☐ 14 – UiT	☐ 15 - IORAS
	☐ 16 - MGO		
Due Date	28 February 2017		
Delivery Date	Version 1: 2 May 2017; Version 2: 30 November 2018		

This project has received funding from the European Union's Horizon 2020 Research & Innovation programme under grant agreement No. 727862.

Table of Contents

EXECUTIVE SUMMARY

Having a thorough model assessment capacity is critical for the APPLICATE project in order to establish the ability of existing models in simulating Arctic weather and climate along with Arctic-midlatitude interactions, provide guidance for APPLICATE model development activities, and to measure the impact and hence success of the APPLICATE project.

This Model Assessment Plan outlines the project's model assessment strategy, making extensive use of the concept of metrics and diagnostics, and utilizing comprehensive sets of observational data. In this context the Plan focuses on evaluating the model's ability to represent

- critical Arctic processes,
- linkages between the Arctic and mid-latitude weather and climate, and
- user-relevant parameters.

The plan also outlines how the metrics and diagnostics will feed into ESMValTool, a tool used by the international research community for evaluating climate models.

Importantly, the Plan also considers the evaluation of numerical weather prediction (NWP) models and seasonal forecasting systems. This assessment will not only provide the basis for enhancing prediction capacity; it will also lead to insight into the origin of model error and thus contributes to providing guidance for the design of the observing system.

Regional Arctic heat budget analyses will be used to assess the ability of models to reproduce key processes in the Arctic and to identify important feedbacks and processes affecting Arctic climate variability and sea ice change.

Towards the end of the project, the potential will be explored for the heat budget and process-based metrics developed within APPLICATE to provide observation-based constraints on the climate models to reduce uncertainty in future projections.

To facilitate effective management of WP1 timelines are given for the different and critical relationships to other WPs are outlined.

The APPLICATE Model Assessment Plan is a "living document" that will be updated regularly. This version 2 (November 2018) is to be considered an update of the original deliverable (May 2017) and includes the progress achieved in the first half of the project and the outcome of the discussion following the first project review (September 2018).

1. INTRODUCTION

1.1. Background and Motivation

To have confidence in climate projections and weather forecasts, it is essential that the models used to make such predictions are capable of capturing key physical processes in the oceans, atmosphere and cryosphere. The aim of WP1 is to develop advanced metrics and diagnostics that will be used to observationally constrain weather and climate models.

The initial focus of WP1 will be on metrics, i.e. the quantitative comparison of a quantity within a model to some reference, for example, an observational data set. Several types of metrics will be developed: process-based metrics to evaluate Arctic climate processes and the linkages between the Arctic and the Northern Hemisphere, user-relevant metrics, and novel sea ice metrics based on observations from YOPP (the Year Of Polar Prediction). User-relevant metrics will be co-developed with users engaged within WP7. To ensure community and user engagement, the metrics developed during APPLICATE will be made widely available through ESMValTool, where applicable.

The metrics developed in WP1 will be used to assess weather and climate models. This includes the existing CMIP5 data, but also the output from the forthcoming CMIP6 activity that will inform the next IPCC assessment report on climate change. The metrics will also be used to assess the ensemble weather forecasts from NWP models.

To encourage the exchange of ideas between the weather and climate modelling communities, a synthesis will be made of how model errors develop across time scales in weather and climate models.

A new generation of process-based heat budgets of the Arctic will be used to investigate the feedbacks and processes that lead to uncertainty in Arctic climate model projections.

WP1 will also explore the potential of observational emergent constraints to reduce the uncertainty in climate model projections for the Arctic and its linkages to the whole Northern Hemisphere.

1.2. Organisation of the plan

The Model Assessment Plan has been developed by the WP1 co-leads (Thomas Jung, AWI, and Len Shaffrey, UREAD) and the APPLICATE project manager (Luisa Cristini, AWI) in conjunction with the WP1 Task Leaders and the institutional researchers participating in WP1. It has been updated in November 2018 taking into account feedback from the first period report; furthermore, progress this update reflects progress made in WP1 since the first submission of the Plan.

The Model Assessment Plan is organised as follows:

Section 2: Model Assessment Strategy. This section describes the wider context for the assessment of weather and climate models and defines the specific terms (metrics, diagnostics, etc.) that are used in the Model Assessment Plan. Section 2 also outlines the conceptual framework for model assessment within APPLICATE.

Section 3: Model Assessment Activities. Section 3 describes the four Model Assessment Activities in WP1 (new targeted metrics; assessment of weather and climate models; heat budgets in the Arctic; emergent constraints). For each Model Assessment Activity, a set of subtasks and associated timeframes are provided.

Section 4: Risks and Interdependencies. This section describes the risks and interdependencies associated with the Model Assessment Plan. Section 4 identifies the key internal interdependencies between WP1 and the other WPs that are essential for the success of the Model Assessment Plan. Section 4 also indentifies external dependencies with key partners, including other H2020 projects and scientific communities.

Section 5: Implementation of the Plan. This section details how the Model Assessment Plan will be implemented, how outcomes will be measured and how progress with the plan will be reported within APPLICATE. Section 5 also outlines how the Model Assessment Plan will be updated and revised throughout the lifetime of APPLICATE, so that the Model Assessment Plan becomes a "living document".

2. MODEL ASSESSMENT STRATEGY

2.1. Motivation

One of the overarching goals of APPLICATE is to improve sub-seasonal to seasonal climate predictions and climate change projections in the Arctic and beyond. To formally detect such improvements and disentangle them from background noise, the development of meaningful performance metrics (e.g., Knutti et al., 2010; Eyring et al., 2016; Flato et al., 2013) simply referred to as "metrics" hereafter, will be a key ingredient to the success of the project. The use of metrics has been pervasive, but also controversial in the history of climate science. Well-chosen metrics are unrivalled tools to make a crisp summary of complex information and to assess climate models or prediction systems, in particular to highlight their major deficiencies. However, simplicity has a price, namely the risk of over-interpretation. Metrics are numbers; numbers are subject to ranking, and rankings almost systematically create an insidious atmosphere of competition between research centers.

The purpose of this document is to lay the foundations of the general strategy that will be followed by the APPLICATE consortium for model and prediction system assessment during the project. More specifically, this document has two goals. First, it aims at proposing unambiguous *definitions* for terms that are commonly used but often loosely defined in the climate and weather communities (or used interchangeably) such as "metric", "diagnostic" or "constraint". Second, it aims at *framing* the development of metrics in APPLICATE by proposing a set of criteria that would make such metrics desirable, attractive and useful for the project.

This plan integrates and synthesizes multiple discussions that took place during the preparation of APPLICATE, during other related projects[1], during APPLICATE's kick-off meeting and first General Assembly in Barcelona as well as during the first period APPLICATE review. As much as we can, we are trying to align with recommendations and definitions from the Intergovernmental Panel on Climate Change (IPCC)'s guidance paper on Assessing and Combining Multi Model Climate Projections (Knutti et al., 2010). While very comprehensive, this document is not entirely fit for the purpose of APPLICATE in which climate prediction is a central theme, and in which novel concepts like 'climate services' are present.

2.2. Wording: Name it!

Agreeing on definitions is a prerequisite for effective communication throughout the project. In the following, climate models, weather models and the corresponding prediction systems under assessment are referred to as the **systems**, while the baselines to which they are

[1] PRIMAVERA (http://www.primavera-h2020.eu), CRESCENDO (https://www.crescendoproject.eu/) among others

compared are termed the **references** (observational products, reanalyses, or even other models).

Diagnostics are quantities derived from geophysical data sets. The definition proposed by Knutti et al. (2010) suggests that diagnostics are exclusively derived from model output; our definition is somewhat larger and also includes observational references and reanalyses. The sea ice extent retrieved from satellite observations of sea ice concentration, the strength of the snow-albedo feedback in a reanalysis or the average eddy kinetic energy of the atmosphere in a coupled climate model over the North Atlantic are all examples of such diagnostics. As such, a diagnostic is a tool to simplify complex information that lives in a high-dimensional physical, temporal, probabilistic space, into something much more easily to digest like maps, time series or histograms.

"In my model, the average 1980-2000 March Arctic sea ice area is 11.73 million km²"

(Diagnostic)

User-relevant diagnostics are a particular type of diagnostics tailored for the ever-growing community of users of climate data such as the insurance sector, governments, the tourism industry and more broadly stakeholders. These diagnostics have generally undergone a high level of processing and tailoring, since they should be usable directly as an input to decision making. In addition, such diagnostics are only disseminated if the quality of the underlying model and prediction system has been thoroughly tested (see "forecast quality metrics" below). By contrast to standard diagnostics, user-relevant diagnostics attempt to characterize the likelihood of well-defined regional climatic events (e.g., probability of experiencing frost in Paris during the next winter) rather than the value of large-scale quantities (e.g. global-mean surface temperature in 2016).

"There is a probability of 93% that the Arctic will not be navigable over the next month of March: in 56 out of 60 members of my forecast system, it is not possible to find a continuous path from the Atlantic to the Pacific along which sea ice concentration and thickness remain below 15% and 0.5cm, respectively"

(User-relevant diagnostic)

Metrics (used interchangeably with performance metrics in this document) are quantitative measures of agreement between a simulated and observed quantity which can be used to assess the performance of individual models (Knutti et al., 2010). Thus, metrics reflect the agreement of a diagnostic from a system with respect to the same diagnostic computed from a reference. More precisely, a metric maps a diagnostic to a single real number, given a reference. Metrics are inherently attached to the notion of "distance" in geometry. Ideally, they should be defined according to a set of axioms too (such as positivity, triangle inequality, symmetry, nullity). Several types of metrics must be distinguished from each other:

- ***Standard error metrics*** are developed in order to check the overall consistency of a model or prediction system with a reference. Standard error metrics are useful: they put pressure on centers to be responsive in addressing obvious model biases, but they also allow for tracking the first-order evolution of model development through time

(Gleckler et al., 2008; Reichler and Kim, 2008; Eyring et al., 2016). Such metrics should be handled by "responsible adults" because they are easily over-interpreted. For instance, a model may simulate a realistic trend in annual-mean, global-mean near-surface air temperature, but thanks to the cancellation of major regional biases. Ideally, standard error metrics should never be computed in isolation (e.g. for one specific variable) but rather be part of an overall assessment process – this would allow an instant visualization of the system's consistency with the reference(s) as a whole.

"The root mean squared error of Arctic sea ice thickness in my model is 1.2 m over 2004-2008, compared to the ICESat sea ice thickness dataset."

(Standard error metric)

- ***Predictability metrics*** provide a quantitative estimation of the predictable content of a system. Predictability metrics are generally derived independently from external references, because the reference used is precisely a slightly different version of the system itself. That is, these metrics result from the comparison of twice the same diagnostic computed from two slightly different versions of the same system. The rate of error growth in global mean temperature between two members of the same model but initialized from slightly different states is an example of such a metric. The e-folding time scale of the autocorrelation function of a given signal (from a model or from observations) can also be considered as a predictability metric, since it is obtained from the comparison (here, correlation) between two slightly different versions of the same diagnostic (here, lagged versions of the signal).

"The spread of my ensemble reaches 95% of the climatological spread after 5 years, giving an approximate bound on predictability for my system".

(Predictability metric)

- ***Forecast quality metrics*** test the ability of a prediction system to re-forecast past events in order to gain confidence about its ability to predict future outcomes. The assessment of **deterministic** forecasts is achieved through the application of classical metrics such as the correlation, the root mean square error or the mean bias between the system's prediction and the reference. Besides, a wide range of metrics has been developed to assess the validity of **probabilistic** forecasts, such as rank histograms (their flatness), Brier skill scores and continuous rank probability skill scores among others. Forecast quality metrics are unique in that they measure the instant correspondence between the system tested and the reference whereas other types of metrics rather focus on the agreement between estimators (means, trends, frequency distributions).

"My system has been able to forecast the observed March winter sea ice extent variations in the Arctic with 87% of explained variance. I'm confident that the prediction for the next month of March will be skillful, and will be superior to simple persistence and climatological forecasts."

(Forecast quality [here, deterministic] metric)

- ***Process-based metrics*** (or process-oriented metrics) aim at evaluating the ability of a system to simulate a particular process, a coupled mechanism or a feedback, based on a physical diagnostic that can in addition be computed from a reference. This class of metrics should help the scientist identifying the reasons behind good or bad model performance by going further than the first-order information offered by standard error metrics. As such, process-based metrics represent a natural extension to standard error and forecast quality metrics (initial tendency errors in medium-range forecasts are good examples of process-based metrics, since they aim at understanding the development of systematic errors in the forecasts). Since the boundary may not always be clear between the meaning and purpose of process-based vs. standard error metrics, the following rule may be kept in mind: standard error metrics measure the ability of a system to simulate physical *states* (regardless of why this is so) while process-based metrics measure the ability of the system to simulate the physical *phenomena* leading to these states, which is a far more constraining requirement.

"A heat budget analysis of my simulation shows Arctic sea ice is area is too low in the Barents Sea due to an excess of meridional heat transport from the Northern Atlantic Ocean. This issue was traced back to the unrealistic parameterization of air-sea fluxes in my model, whereby turbulent heat fluxes are overestimated by a factor of 3 implying too much heat absorption by the ocean in the model."

(Process-based/oriented metric)

A **constraint** is the application of a metric to an ensemble of models displaying relationships between two diagnostics, one of which can be observed. Since this relationship "emerges" from the ensemble, the wording **emergent constraint** is often used (e.g., Collins et al., 2012). As an example, Hall and Qu (2006) find a relationship between the strength of the snow albedo feedback computed over a season and the strength of the same feedback estimated over this century, in the CMIP3 ensemble. They use the first diagnostic (seasonal albedo feedback) as a constraint for the second one (century albedo feedback) using observations available. Naturally, the use of emergent constrained should be accompanied by solid physical understanding to rule out the possibility of spurious correlations.

"In a hierarchy of climate models, meridional oceanic heat transport in the North Atlantic over 1980-2000 is negatively correlated to the loss of Arctic sea ice volume between 2000 and 2050. This relationship is not accidental: it can be explained using physical arguments. Based on the estimated oceanic transport from observations, I find that the Arctic may lose between 11 and 15 thousands of km³ of ice on annual-mean between 2000 and 2050.

(Emergent constraint)

Finally, a **diagnosis** is an integrated statement about a model or a forecast system evaluated for a certain purpose. It involves the use of diagnostics and different metrics, together with prior knowledge about the system itself and its underlying physics. A diagnosis aims at resolving problems by looking at causes rather than symptoms. Unlike diagnostics, diagnoses are by definition not runnable by computers: they require expertise, exchanges through discussions, and synthetic thinking.

2.3. The CRISTO framework for metrics in APPLICATE

Since no diagnostic and by extension no metric is all-purpose, the question "What is the best metric" must be rephrased by "What criteria should good metrics fulfill?" We propose a set of six criteria that an ideal metric or set of metrics should meet. These guidelines are summarized by the acronym "CRISTO" for Completeness, Rationale, Interpretability, Stability, Transparency and Observability.

1) **Completeness**. By construction, a single metric cannot verify the validity of a system exhaustively[2]. However, a well-chosen set of metrics covering various variables on different time scales and in different regions may provide a good idea as whether the system is in overall agreement with a set of references or not. As much as possible, such an ensemble of metrics should be as *complete* as possible, meaning that the metrics should together cover all relevant aspects for which the system is to be evaluated. Having a *minimal* number of metrics to achieve this goal is also a desirable property. Hence, an ideal set of metrics should have the same properties as a 'basis' in linear algebra: it should be complete while and individual metrics should be as orthogonal from each other as possible (i.e. not redundant).
2) **Rationale**. A metric should always be defined with a clear scope in mind, and according to a scientific question clearly stated *a priori*. That is, the design of a metric should be the last step in scientific reasoning and should allow to test ultimately if the initial hypothesis (of model improvement for instance) is verified or not. Working the other way around (i.e., applying the metrics first and then formulating a scientific statement) exposes to the risk that scientific conclusions are adjusted – consciously or not – to match the results obtained. Indeed, there are so many different ways to measure the skill of a system that it is often possible to highlight at least one aspect in which it has improved.
3) **Interpretability**. Metrics are numbers and numbers abstract objects. Apart from the person who designed and computed the metric, it is likely that virtually no one will have a good understanding of what is exactly meant by that particular metric. Therefore, a good metric should always be accompanied with supporting information: a short description, a figure or an animation. For instance, if a correlation is used to underline the skill of a system to forecast the NAO, the individual time series of the forecasted and observed NAO, as well as the scatter plot between the two series should at least be provided as side information – this could allow to show the presence of outliers or nonlinearities, for example, and question the meaning of the correlation displayed as a proof for skill.
4) **Stability**. A good metric should be stable with respect to internal variability and interannual variability in the system assessed. In addition, it shouldn't be affected too much by uncertainty in the reference. That is, the conclusions should be insensitive to the time period chosen, to the observational product used, or to the member picked from the model. This is far from obvious, but if this is the case, then targeted observational campaigns such as the ones carried out during the Year of Polar Prediction (YOPP) represent invaluable opportunities to conduct efficient and meaningful model evaluation. In any case, a good metric should ideally be communicated in a probabilistic way (i.e. as a random variable with a PDF) rather than in a deterministic way (i.e. as a fixed number), in order to remember that metrics

[2] Whether a system is verifiable *at all* is part of another philosophical debate that is out of the scope of this note.

themselves are uncertain due to the imperfect experimental conditions in which they are developed.

5) **Transparency**. A good metric should be fully reproducible. It should be coded in an open-source language and easy to share (ideally through a version control software such as Git or SVN), so that anybody is free to verify the steps leading to the final result. This is the best way to respond to criticisms that the use of selected metrics will inevitably raise, especially when it comes to model ranking and selection. Furthermore, by making the process of evaluation fully transparent, anyone willing to propose alternative metrics will have the possibility to do so. Moving to community tools such as the ESMValTool[3] seems the most obvious way forward to maximize transparency in the design of future metrics (for CMIP6 among others).
6) **Observability**. A good metric should be derived from diagnostics that are easily observable. Sometimes, this is not possible or very difficult (think for example of the heat budget of sea ice). In this case, a diagnostic can still be useful in explaining model-to-model differences, but tracking down the root causes of model biases might be more difficult.

Working within the "CRISTO" framework described above should not be seen as an obligation but rather as a recommendation. In fact, it is virtually impossible to find examples of metrics that fulfill all six points at once. The guidelines presented here allow to make sure that the assessment of models, reanalyses and prediction systems is conducted in APPLICATE follows strict scientific standards. At the same time, following these recommendations will minimize the risk of over-interpretation since all contextual and technical elements would be provided to appreciate the purpose and limitations of the metrics under consideration.

3. MODEL ASSESSMENT ACTIVITIES

3.1. New targeted metrics for model assessment

3.1.1. Earth System Model Evaluation Tool (ESMValTool)

In Tasks 1.2 of WP1, entitled "Weather and climate model evaluation", ESMValTool will be used extensively for model assessment. ESMValTool (Earth System Model eValuation Tool) is a community tool for evaluating metrics and diagnostics from Earth System Models (ESMs). ESMValTool allows for routine comparisons of single or multiple models, either against predecessor versions and/or observations. ESMValTool is a community effort open to both users and developers encouraging open exchange of diagnostic source code and evaluation results. In this regard, the uptake and development of ESMValTool within APPLICATE is aligned with model assessment goals of APPLICATE and the project's commitment to open data.

Selected new metrics and diagnostics relevant for the Arctic will be incorporated in ESMValTool and made available to the APPLICATE consortium and the international science community. Initially, it was planned to incorporate most of the metrics and diagnostics, developed by APPLICATE partners in ESMValTool. However, it turned that immediate incorporation of new metrics and diagnostics, without thorough understanding of their strengths and short-coming, would not be advisable. For APPLICATE this will be mean that fewer, but well understood new metrics and diagnostics will be incorporated in ESMValTool at a later stage of the project than originally planned.

At the kick-off meeting, it was decided that rather than having one partner being responsible for ESMValTool, each of the partners involved in Task 1.2 should develop expertise in using

[3] https://www.esmvaltool.org

ESMValTool. By choosing this approach, APPLICATE will contribute to increasing the size of the research community that is capable of advancing ESMValTool.

In order to enable consortium members to more effectively use and develop ESMValTool, two training activities were carried out: An online tutorial held in early summer 2017 targeting beginners; and a second tutorial, including a Q/A session at the first next APPLICATE General Assembly.

3.1.2. Development of process-based metrics for the Arctic and implementation in ESMValTool (**AWI**, SU, CNRS-GAME, UCL, UiB, UNI Research, IORAS, MGO) (M1-M12)

APPLICATE will develop a series of metrics to enable the assessment of weather and climate models in their ability to represent critical processes in the Arctic, as well as the ability to capture observed linkages with lower latitudes. Emphasis will be placed on those processes being targeted in APPLICATE's model developments efforts (WP2). Central to this effort is also the gathering of relevant high-quality observational data sets, both from existing (e.g. Obs4MIPs- and reanalysed) and new sources (e.g. Earth observations). Key-metrics will be made available through ESMValTool. Given that ESMValTools requires the availability of high-quality observational data, "dissemination" of observational reference data sets will be made available through ESMValTool.

Table 1: *Process-based* metrics and diagnostics for the Arctic which will be implemented in ESMValTool.

Atmosphere	**Snow**	**Sea-ice**	**Ocean**
Atmospheric boundary layer	Snow cover	Dynamics	Subset of metrics used in CORE-II efforts
Clouds	Snow on sea ice	Thermodynamics	Arctic ocean circulation
Circulation		Sea ice thickness	Arctic ocean water mass characteristics
Storm track activity			Arctic freshwater budget
			Exchanges of heat and mass through Arctic gateways

Table 2: Timeline for development of process-based metrics for the Arctic and implementation in ESMValTool

Due date (project month)	**Activity**
8	Installation of ESMValTool at partner institutions
8	ESMValTool training session for beginners
9	Tests of ESMValTool with existing data
9	Finalise list of process-based metrics together with WP2
10	Develop process-based metrics available through ESMValTool
12	First set of process-based metrics available through ESMValTool
14	ESMValTool Question and Answer Session at APPLICATE General Assembly

3.1.3. Co-Development of user-relevant impact metrics and implementation in ESMValTool (**UREAD**, AWI, BSC, IORAS) (M1-M12)

APPLICATE will co-develop a series of metrics together with stakeholders, who are interested in the impacts that come with Arctic climate change and its effect on Northern Hemisphere weather and climate. Initial conversations that APPLICATE partners have held with stakeholders have revealed the list of parameters given in Table 3. This table will be further developed in conjunction with WP7 and relevant stakeholders. Key metrics will be made available through ESMValTool.

Table 3: User-relevant impact metrics which will be implemented in ESMValTool.

Atmosphere	Snow	Sea-ice	Ocean
Severity and frequency of strong winds associated with storms as a hazard to shipping, fishing vessels and coastal communities	Snow cover	Sea ice free regions for shipping	Ocean temperature for fisheries
Winds for wind farm operators and transmission systems		Location of the ice edge	
Temperature and precipitation for food security in the Arctic and beyond			

Table 4: Timeline for development of user-relevant impact metrics and implementation in ESMValTool

Due date (project month)	Activity
6	Installation of ESMValTool at partner institutions
8	Tests of ESMValTool with existing data
8	Finalise list of user-relevant metrics together with WP7
10	Develop user-relevant metrics available through ESMValTool
12	First set of user-relevant metrics available through ESMValTool

3.1.4. Development of metrics that describe linkages in atmosphere and ocean and implementation in ESMValTool (**CNRS-GAME**, AWI, CERFACS, UREAD, IORAS) (M1-M24)

APPLICATE will develop metrics that can be used to quantify and assess linkages between the Arctic and the Northern Hemisphere atmosphere and ocean. Key-metrics developed in this task will be made available through ESMValTool

Table 5: Mid-latitude linkage metrics which will be implemented in ESMValTool.

Atmosphere	Ocean
Interannual lead-lag relationships between Arctic sea ice/Siberian snow cover and the AO/NAO and ENSO	Mass, heat and fresh water fluxes through Fram Strait, the Barents Sea and the Canadian Arctic Archipelago.
Flow-dependence of atmospheric Arctic-mid-latitude linkages	
Equator-to-pole temperature gradient metrics	

Table 6: Timeline for development of mid-latitude linkage metrics and implementation in ESMValTool

Due date (project month)	Activity
8	Installation of ESMValTool at partner institutions
12	Tests of ESMValTool with existing data
14	Finalise list of metrics that describe linkages in atmosphere and ocean
20	Develop metrics that describe linkages in atmosphere and ocean including observational datasets
24	First set of metrics that describe linkages in atmosphere and ocean available through ESMValTool

3.1.5. Development of novel sea ice metrics from YOPP special observing periods and implementation in ESMValTool (UCL) (M1-M36)

YOPP will include several intensive observing periods (IOPs) during which comprehensive observational datasets will be generated. APPLICATE will work with YOPP observational groups to develop new sea ice metrics focused on specific processes and/or feedbacks which will be robustly sampled even though the observing periods are relatively short. These novel metrics will be used for model assessment in APPLICATE. These metrics will also be disseminated to the groups participating in the YOPP IOPs, thereby further establishing links between modelling and observational experts. Metrics will be made available through ESMValTool.

Table 7: Novel sea ice metrics which will be implemented in ESMValTool.

Sea ice
Heat conduction through sea ice
Sea ice-albedo feedback

Table 8: Timeline for development of novel sea ice metrics and implementation in ESMValTool

Due date (project month)	Activity
6	Installation of ESMValTool at UCL
8	Tests of ESMValTool with existing data
16	Finalise list of novel sea ice metrics
24	Develop novel sea ice metrics including observational datasets
36	First set of novel sea ice metrics available through ESMValTool

3.2. Assessment of weather and climate prediction models

APPLICATE will assess the ability of weather and climate models to represent key processes in the Arctic, linkages between the Arctic and Northern Hemisphere, and user-relevant metrics. The assessment will serve as the baseline from which the model developments carried out in WP2 of APPLICATE will be evaluated.

3.2.1. Assessment of CMIP5 and CMIP6 climate models (**UREAD**, AWI, MGO) (M9-M30)

APPLICATE will use the metrics outlined in Section 2.1 for assessing CMIP5 HISTORICAL simulations and the CMIP5, RCP2.6, RCP4.5 RCP6.0 and RCP8.0 climate change simulations. This assessment will determine i) the systematic errors in the CMIP5 models and ii) the ensemble mean and inter-model spread in CMIP5 climate projections. Particular attention will be paid to assessing the sampling uncertainties in metrics that arise from the internal variability inherent in observations and climate models.

As they become available during 2018 and 2019, the metrics will be used to assess the CMIP6 climate model simulations, especially those carried out by the APPLICATE partners. The assessment of CMIP6 models will also identify any potential reductions in systematic errors between CMIP5 and CMIP6 and potential changes in climate projections. Furthermore, this assessment will also provide the baseline assessment for model developments in WP2 and inform the numerical experiments in WP3.

Table 9: Timeline for assessment of CMIP5 and CMIP6 climate models

Due date (project month)	Activity
16	Assess CMIP5 HISTORICAL simulations: Process-based and user-relevant metrics
16	Assess CMIP5, RCP2.6, RCP4.5 RCP6.0 and RCP8.0 climate change simulations
20	Assess first CMIP6 runs (HISTORICAL and/or HiResMIP) from APPLICATE partners: Process-based and user-relevant metrics
24	Assess full CMIP6 data (HISTORICAL and/or HiResMIP): Critical processes and user-relevant parameters
26	Assess first CMIP6 runs (HISTORICAL and/or HiResMIP) from APPLICATE partners: Linkages between the Arctic and midlatitudes
30	Assess full CMIP6 data (HISTORICAL and/or HiResMIP): Linkages between the Arctic and midlatitudes

3.2.2. Assessment of NWP systems (**ECMWF**) (M12-M48)

APPLICATE will establish and test a diagnostic framework that will be applied to short-to-medium range predictions and initial conditions to establish sources of model error (guidance for WP2) and the impact of observational data in WP4. Furthermore, this task will contribute to the revision of atmosphere and snow model components in WP2 and guide observing system experiments in WP4. Novel diagnostics targeting the coupled surface-atmosphere-snow-sea ice system will be developed and applied to identify key sensitivities in coupled models and key sources of model error. These diagnostics will also be used to support the model development in single-column mode.

Diagnostics linking the contributions from individual physical processes to model tendencies and analysis increments in the atmosphere will be developed. These diagnostics will allow model error to be traced back to individual processes represented in short-range forecasts. Furthermore, the statistics of analysis increments will allow for an evaluation of the impact of observations on the analysis.

The statistics from ensemble data assimilation can also be used to assess model and observation contributions to ensemble spread in NWP systems. This provides guidance for model error formulation in ensemble systems, but also insight into the observational impact in the analysis. This will be further exploited in WP4. APPLICATE will also support recommendations for operational monitoring and evaluation capabilities dedicated to polar

requirements. A demonstration of such monitoring will be introduced with a focus on surface radiation, cloud and snow observation networks and satellite retrievals.

Table 10: Timeline for assessment of NWP systems

Due date (project month)	Activity
12	Establish diagnostic framework
20	Contribute to the revision of the atmospheric and snow model components
20	Provide recommendations for observing system experiments in WP4
24	Develop novel diagnostics targeting the coupled surface-atmosphere-snow-sea ice system and apply for the identification of key sensitivities in coupled models and key sources of model error.
36	Develop and apply tendency and analysis increment diagnostics in the Arctic
36	Evaluate the impact of observations on the analysis
36	Assess observation and model contributions to ensemble spread in NWP systems
48	Make recommendations for operational monitoring and evaluation capabilities dedicated to polar requirements. Provide demonstration for such monitoring with a focus on surface radiation, cloud and snow observation networks and satellite retrievals

3.2.3. Synthesis: Growth of model error across time scales (**AWI**) (M24-M36)

The insights gained from the assessment of weather and climate models will be synthesized to improve our understanding of the processes that lead to common model errors in both weather and climate models. To this end, the YOPP Analysis and Forecast Data Set (WP6) will also be exploited. This seamless approach to model error diagnosis will enable APPLICATE to identify error in climate models that are determined by processes occurring on relatively short time scales from hourly to weekly. This in turn will inform the model development activities in WP2. Understanding the commonalities in error growth will also help foster the exchange of ideas for model development between the weather and climate modelling communities.

Table 11: Timeline for synthesis

Due date (project month)	Activity
24	Gather model data
30	Evaluate error growth across time scales
36	Provide synthesis report

3.3. Assessment of Arctic heat budget in climate models

We will assess the ability of the climate models used in APPLICATE to represent the seasonal cycle and long-term trends in the heat budget of the Arctic – including atmosphere, ocean, sea ice and snow components. Building on the approach of Keen et al. (2013), the assessment of the Arctic heat budget will identify important feedbacks and processes that govern Arctic climate variability and change. Coordinated analysis of the CESM climate model will also be performed at NCAR in the US. This work will also contribute to the IPCC through SIMIP (Sea Ice Model Intercomparison Project).

A detailed assessment of the heat budget of the Arctic Ocean will also be performed. It will focus on the links between changes in oceanic heat transport into the Arctic and Arctic sea ice. The vertical mixing processes that redistribute heat within the Arctic Ocean, and thus lead to impacts on sea ice, will also be evaluated. The ability of the APPLICATE models to capture the observed heat budget of the Arctic will be evaluated using the metrics developed in WP1. A synthesis of the two heat budget approaches will be made.

This activity is led by the MET OFFICE.

Table 12: Timeline for assessment of Arctic heat budget in climate models

Due date (project month)	Activity
40	Development and implementation of sea ice volume/energy budget analysis
40	Development and implementation of ocean heat budget analysis
48	Synthesis from the ocean and sea ice heat budget methods

3.4. Assessment of the utility of observational emergent constraints in reducing the uncertainty of CMIP5 and CMIP6 climate change projections in the Arctic and mid-latitudes

APPLICATE will explore the potential of the metrics and analysis in Tasks 1.2, 1.3 and 1.4 of WP1 to provide observational emergent constraints on climate model projections. Emergent constraints are metrics which show strong relationships between the biases in historical climate model simulations and the sensitivity of climate projections. An example of an emergent constraint is the relationship found between biases in Arctic sea ice cover and future trends in CMIP3 climate models (Boé et al, 2009).

Emergent constraints potentially enable the uncertainty in climate model projections to be reduced, since projections from climate models with smaller biases should be more plausible.

In APPLICATE, emergent constraints in the Arctic will be investigated using biases in the seasonal cycle, past trends, interannual variability, and their relationship with Arctic climate change. Emergent constraints will also be investigated in the context of links between the Arctic and Northern Hemisphere atmospheric and oceanic circulation, e.g. through changes in Arctic warming, in the equator-to-pole temperature gradient and in subsequent impacts on mid-latitude atmospheric circulation. A synthesis of the results on emergent constraints will be made.

This activity is led by the MET OFFICE.

Table 13: Timeline for assessment of the utility of observational emergent constraints in reducing the uncertainty of CMIP5 and CMIP6 climate change projections in the Arctic and mid-latitudes

Due date (project month)	Activity
16	Summary of planned work on emergent constraints
48	Synthesis of results of emergent constraints

4. RISKS AND INTERDEPENDENCIES

4.1. Internal within the project

Assessing the Risk and Interdependencies of the Model Assessment are essential for the successful implementation of WP1. In this section possible risks and critical interdependencies within APPLICATE (e.g. between WPs) are outlined. Table 14 summarizes important output from WP1 to other WPs and vice versa. A more comprehensive risk register is given in the project Risk Management Plan (deliverable 9.3).

Table 14: Internal interdependencies between WP1 and other WPs

WP1 provides to	**Output**
WP2	▪ Model evaluation framework ▪ Baseline assessment of process-based metrics and Arctic linkages in climate and weather models ▪ Tools and metrics to assess improvement in climate and weather models
WP3	▪ Model evaluation framework ▪ Guidance on how to design physically plausible numerical experiments addressing atmospheric and ocean linkages
WP4	▪ Model and data assimilation evaluation framework ▪ Key metrics for assessment of model forecast skills
WP5	▪ Model evaluation framework ▪ Key metrics for assessment of model forecast skills ▪ Scientific basis for quantifying impacts ▪ Advice on how emergent constraints in the Arctic can reduce uncertainty of climate change projections
WP7	Assessment of the ability of weather and climate models in representing and predicting user-relevant parameters.

WP1 receives from	**Input**
WP2	▪ Advice on the design of metrics and diagnostics targeting those processes that will be enhanced as part of APPLICATE's model development
WP7	▪ Engagement with stakeholders to co-develop the list of user relevant metrics

4.2. External relationships with other partners

In addition to internal risks and interdependencies, it is essential to consider relationships with external partners. This include other EU projects (where relationship will also be considered in

WP8) and with relevant scientific communities. An in-depth dialogue with other EU projects on model assessment will initiated at a joint meeting held from 23-24 May 2017 in Brussels.

Table 15: Interdependencies between APPLICATE and external partners

External partners	Collaboration
CRESCENDO	▪ Development of ESMValTool ▪ Sharing of metrics and diagnostics
PRIMAVERA	▪ Development of ESMValTool ▪ Sharing of metrics and diagnostics
Blue-Action	▪ Sharing of metrics and diagnostics
CORE-II/OMIP community	▪ Metrics for assessment of ocean components ▪ Advice on gridding
SIMIP community	▪ Metrics for assessment of sea ice models
ESMValTool Community	▪ Access to ESMValTool source ▪ Guidance on training, use and development

5. IMPLEMENTATION OF THE PLAN

The detailed timelines for each of the tasks in the Model Assessment plan is given in Section 2. Each of the tasks is assigned to a lead organisation, which is responsible for reporting progress. Progress on the tasks in the Model Assessment plan is reported to WP1 leaders (Thomas Jung, AWI and Len Shaffrey, UREAD). The WP1 Leaders in turn report on WP1 progress to the Executive Board.

The Model Assessment Plan will be updated regularly (at least once per year) and updates will be presented to and approved by the Executive Board. Further updates to the Plan are foreseen for: February 2019 and February 2020.

6. REFERENCES

Collins, M., R. E. Chandler, P. M. Cox, J. M. Huthnance, J. Rougier and D. B. Stephenson, 2012, Quantifying future climate change, Nature Climate Change 2, 403–409

Eyring, V. *et al.*, 2016: ESMValTool (v1.0) – a community diagnostic and performance metrics tool for routine evaluation of Earth system models in CMIP, *Geosci. Model Dev.* 9, 1747-1802

Flato, G., J. Marotzke, B. Abiodun, P. Braconnot, S.C. Chou, W. Collins, P. Cox, F. Driouech, S. Emori, V. Eyring, C. Forest, P. Gleckler, E. Guilyardi, C. Jakob, V. Kattsov, C. Reason and M. Rummukainen, 2013: Evaluation of Climate Models. In: Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change [Stocker, T.F., D. Qin, G.-K. Plattner, M. Tignor, S.K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex and P.M. Midgley (eds.)]. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.

Gleckler, P. J., K. E. Taylor, and C. Doutriaux, 2008: Performance metrics for climate models, *J. Geophys. Res.* 113, D06104

Hall, A. and Qu, 2006: Using the current seasonal cycle to constrain snow albedo feedback in future climate change, 33, L03502, doi:10.1029/2005GL025127

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

Knutti, R., G. Abramowitz, M. Collins, V. Eyring, P.J. Gleckler, B. Hewitson, and L. Mearns, 2010: Good Practice Guidance Paper on Assessing and Combining Multi Model Climate Projections. In: Meeting Report of the Intergovernmental Panel on Climate Change Expert Meeting on Assessing and Combining Multi Model Climate Projections [Stocker, T.F., D. Qin, G.-K. Plattner, M. Tignor, and P.M. Midgley (eds.)]. IPCC Working Group I Technical Support Unit, University of Bern, Bern, Switzerland.

Reichler, T. and J. Kim, 2008: How well do coupled models simulate today's climate? Bull. Am. Met. Soc. doi:10.1175/BAMS-89-3-303