APPLICATE Model Assessment Strategy

François Massonnet and Thomas Jung

20th of April 2017

This note is an excerpt of the Model Assessment Plan of the APPLICATE project.

APPLICATE is a four-year Horizon 2020 project involving 16 partners from universities, research centers and operational centers. Its aim is to enhance medium-range and climate predictions capabilities in the Arctic but also to determine the influence of Arctic climate change on lower latitudes.

More information on the project: www.applicate.eu

Motivation

One of the overarching goals of APPLICATE is to improve sub-seasonal to seasonal climate predictions 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" hereinafter, 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 note integrates and synthesizes multiple discussions that took place during the preparation of APPLICATE, during other related projects[1] and during APPLICATE's kick-off meeting. 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 always fully fit for the purpose of APPLICATE in which climate prediction is a central theme, and in which novel concepts like 'climate services' are present.

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

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 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 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 km^3 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.

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

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

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 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] https://www.esmvaltool.org

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

Acknowledgments. We thank all APPLICATE partners for fruitful discussions and useful feedback on the document. In particular, we thank Francisco Doblas-Reyes, Gilbert Brunet, Helge Goessling, Veronika Eyring and Wei Cheng for their useful input.

Feedback, questions and comments are welcome:

francois.massonnet@uclouvain.be and thomas.jung@awi.de