SPECIAL PROJECT PROGRESS REPORT

All the following mandatory information needs to be provided. The length should *reflect the complexity and duration* of the project.

Reporting year	2020
Project Title:	Data-driven calibration of stochastic parametrization of IFS using approximate Bayesian computation
Computer Project Account:	spgbdutt
Principal Investigator(s):	Dr. Ritabrata Dutta
Affiliation:	Department of Statistics, Warwick University, UK.

Name of ECMWF scientist(s) collaborating to the project Dr. Nils Wedi and Dr. Peter Dueben, ECMWF.

Start date of the project: 2020

Expected end date: 2022

Computer resources allocated/used for the current year and the previous one

Please answer for all project resources

		Previous year		Current year	
		Allocated	Used	Allocated	Used
High Performance Computing Facility	(units)	NA	NA	6000000	0
Data storage capacity	(Gbytes)	NA	NA	6400	0

Summary of project objectives (10 lines max)

The aim of the project is to develop Approximate Bayesian Computation (ABC) techniques suitable to be applied in the setting of ensemble NWP models. Specifically, this will allow to perform Bayesian inference of parametrization parameters from observations. Bayesian inference, opposed to frequentist inferences, provides a better way to quantify uncertainty starting from previous knowledge. The ABC paradigm allows moreover to perform inference tasks on highly complex models, only relying on the capacity to simulate model outputs for some parameter values. Achieving this successfully in the setting of NWP models requires the use of state-of-the-art ABC models, in order to reduce the computational burden, together with the development of discrepancy measures that exploit ensemble models information. The ultimate goal of the project is therefore combining these components in a pipeline that can be easily deployed in the specific case of NWP.

Summary of problems encountered (10 lines max)

Due to some miscommunication, the users didn't receive the token giving access to the ecgate computing facilities, for the first 3-4 months of the project duration. After getting access, we did not yet manage to install some crucial Python3 packages on both the ecgate server and cca supercomputer, as some of them are not readily available for the Linux versions ecgate runs, and as another library requires an MPI implementation which does not seem to be available on cca. We expect these problems to be overcome in the near future as we are in touch with servicedesk. This may explain our not using the computational quota at all until now, but we hope to utilize all of this year's quota as soon as we get these issues fixed.

Finally, Lorenzo Pacchiardi (PhD student in Department of Statistics, Oxford) collaborating in this project was supposed to attend the training week on IFS in ECMWF, which was cancelled due to the coronavirus pandemic. This caused some delay in understanding the details of IFS.

Summary of plans for the continuation of the project (10 lines max)

In the near future, we will be developing approaches for calibration of parametrization by using the Lorenz95 model as a benchmark, specially focusing on the methodological challenge of extending ABC to the setting of ensemble models, in particular using skill scores, climatological skill measures and calibration measures as discrepancy measures in ABC (see Summary of Results for more details). After testing these approaches there, we will move on to the IFS simulator, and use data from the Concordia dataset in order to attempt parameter tuning for some parametrization parameters describing convection phenomena.

List of publications/reports from the project with complete references

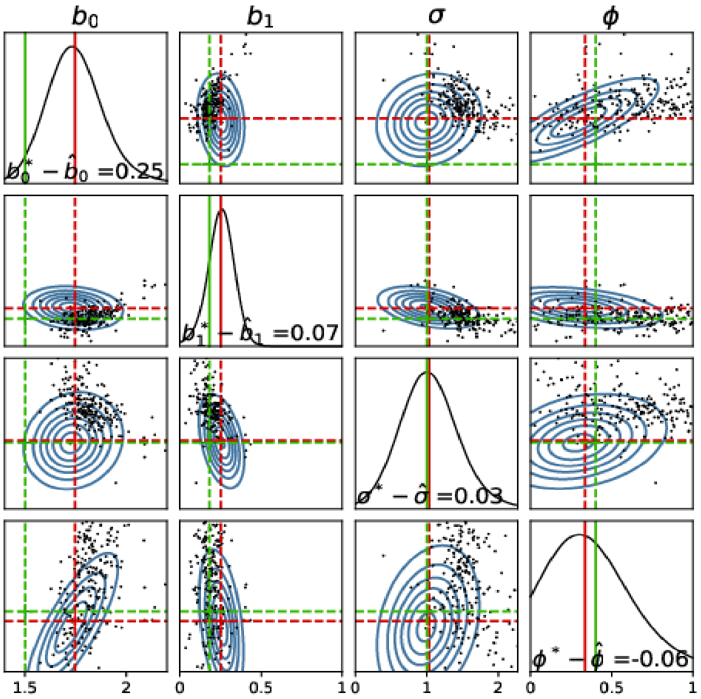
Poster presentation at the Machine Learning for Weather and Climate Modelling in Oxford, in September 2019 (http://users.ox.ac.uk/~phys0895/mlwc2019/Programme.html).

Pacchiardi, L., Künzli, P., Schoengens, M., Chopard, B., & Dutta, R. (2020). Distance-learning For Approximate Bayesian Computation To Model a Volcanic Eruption. *Sankhya B*, 1-30. https://link.springer.com/content/pdf/10.1007/s13571-019-00208-8.pdf

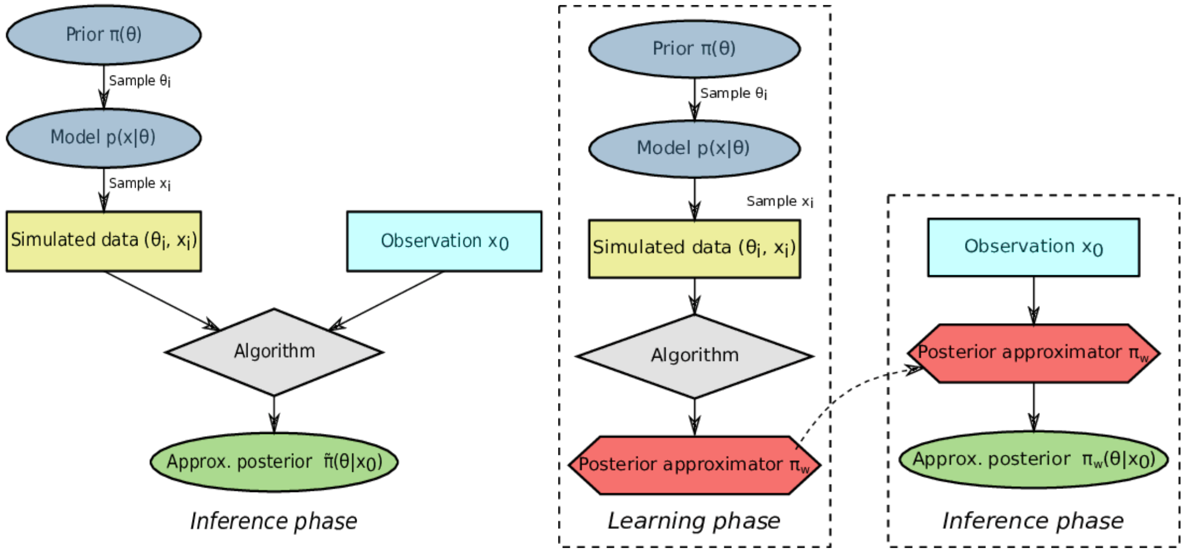
Summary of results

During these first months of the project, the focus has been on developing some methodological techniques for Approximate Bayesian Computation (ABC) which could be applied to the considered application. Also, we started studying the Lorenz95 model before moving to IFS, as a toy model on which to try out the developed techniques.

In specific, the Lorenz95 model considers the evolution of two kinds of variables (“slow” and “fast”) which are coupled by some ordinary differential equations. It is widely used as a testbench for parametrization studies. In particular, we have considered a specific parametrization form and have been able to prove the adequacy of ABC for retrieving parameter values from an observation. In doing this, we have also exploited recent techniques in ABC which employ neural networks in order to reduce the dimensionality of the data, thus drastically reducing the computational burden at the expense of a very mild loss of information. Moreover, a neural network architecture specific for time-series data has been used. The results of this work have been presented in a poster at the Machine Learning for Weather and Climate Modelling in Oxford, in September 2019. The following image shows results of the inference procedure as described above; our approach is able to return a probability distribution over the parameter values, which is marginalized over each parameter in the diagonals, and shown as paired contour plots in the off-diagonal panels. The obtained posterior distribution is concentrated around the true value of the parameter (green line) for all but one parameter value, with the red line representing posterior mean.



The methodological work we have carried out involved firstly the development of a new way to extract information of the data from neural networks using summary statistics by applying distance-learning technique. In a second moment, we have been developing a technique to use neural networks in order to perform *amortized* inference for likelihood-free models (as the Lorenz95 or IFS), where by *amortized* we mean that after a first, computationally expensive, training phase, inferences with the same model on many different observations can be carried out cheaply. This work is still in progress, carried out by Lorenzo Pacchiardi, a PhD student in the Department of Statistics, Oxford.



Case-based inference (left) vs. Amortized inference (right).

Finally, we have started investigating the application of ABC to the setting of ensemble models. In ABC, a discrepancy measure between the output of the model and the observation is needed in order to bypass the absence of a likelihood function and be able to perform (approximate) inference. However, in the case of ensemble models, it is not straightforward to compare the output of the model with the observation (as the output of the model is made up of an ensemble of observations). Therefore, we are starting to investigate the use of ensemble skill scores, climatological skill measures and calibration measures of the prediction in order to perform and evaluate inference in this setting. This will likely lead to new methodological advances which will allow the application of ABC techniques to the NWP setting.