AQUA Extended: Evaluation Framework Bridging DestinE Simulations and ML Climate Emulator

Jose González-Abad[2*], Fernando Iglesias-Suarez[1], Sergio Portilla[1], Marvin Axness-Ferrando[3], Miguel Castrillo[3], Francisco Doblas-Reyes[3,4], Amanda Duarte[3], Aina Gaya-Àvila[3], Hernan Andres Gonzalez Gongora[3], Jose Manuel Gutierrez Llorente[2], Christian Lessig[5], Sebastian Milinski[5], Ankit Patnala[6], Alejandro Peraza[3], Antonio Perez[1], Daniel San Martin Segura[1], Jakob Schloer[5], Martin Schultz[6]

[1]Predictia Intelligent Data Solutions (Predictia); [2]Consejo Superior de Investigaciones Científicas (CSIC); [3]Barcelona Supercomputing Center (BSC); [4]ICREA; [5]European Centre for Medium-Range Weather Forecasts (ECMWF); [6]ForschungsZentrum Jülich (FZJ); ***gonzabad@ifca.unican.es**

1. Introduction

This framework extends AQUA (DestinE's evaluation tool for Climate Digital Twins) to bridge the gap between physics-based and machine learning (ML) climate simulations. Built on top of the **AQUA diagnostic library**, the framework allows reproducible evaluations across HPC platforms like MareNostrum5. It integrates both standard ML metrics (e.g., RMSE, ACC) and climate-specific diagnostics (e.g., ENSO, NAO, MJO).

2. AQUA

AQUA (Climate DT Applications for QUality Assessment) is a model evaluation framework designed for **running diagnostics on high-resolution climate models**, specifically for Climate DT climate simulations being part of Destination Earth activity. The package provides a flexible and efficient framework to process and analyze large volumes of climate data. With its modular design, AQUA offers seamless integration of core functions and a wide range of diagnostic tools that can be run in parallel.

AQUA efficiently handles high-resolution climate data using **catalogues**, **readers**, **fixers**, and **regridders**. It ensures **consistent data access**, **metadata standardisation**, and **regridding**.

```
sources:
  CESM2_r1i1p1f1:
    description: Earth system model: CESM2 (r1i1p1f1)
    driver: yaml_file_cat
    args:
      path: "{{CATALOG_DIR}}/CESM2_piControl_r1i1p1f1/main.yaml"
  emulator_inference:
    description: Machine-learning based climate emulator (different configurations)
    driver: yaml_file_cat
    args:
      path: "{{CATALOG_DIR}}/emulator_inference/main.yaml"
```

Example of a catalogue entry.

```
from aqua import Reader
reader = Reader(catalog='CESM2', model='emulator_inference', exp='historical', source='Experiment_4.3a',
                fix=True)
data = reader.retrieve()
```

Example of data loading with AQUA.

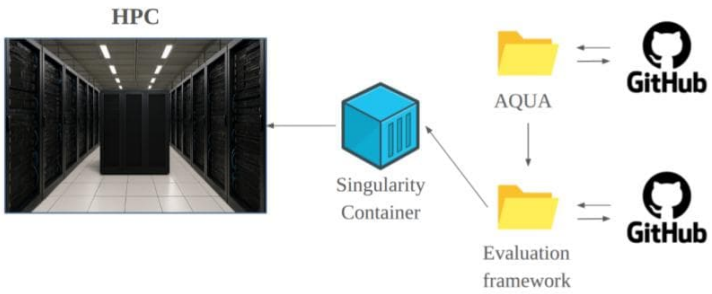
```
plugins:
  source:
    - module: intake_xarray

sources:

  Experiment_4.3a:
    description: Simulations from the Experiment 4.3a.
                 *** Configuration description ***
    driver: netcdf
    args:
        chunks:
            time: 1
        urlpath:
        - <data-path>

    metadata:
      fixer_name: <fixer-name>

  Experiment_5.1c:
    description: Simulations from the Experiment 5.1c.
                 *** Configuration description ***
    driver: netcdf
    args:
        chunks:
            time: 1
        urlpath:
        - <data-path>

    metadata:
      fixer_name: <fixer-name>
```

Example of the definition of various sources.

3. AQUA Diagnostics

AQUA diagnostics are **modular, self-contained components** that evaluate specific aspects of model output. This design allows easy extension and integration with AQUA's data pipeline. Several diagnostics are already available, including tools for major variability modes (ENSO, MJO, NAO), climatological trends, and time series analysis—enabling robust, process-oriented evaluation of emulator performance.

4. Evaluation Framework



High-level structure of the framework and its deployment on HPC systems.

The **evaluation framework** provides all necessary tools to perform end-to-end validation of ML-based emulator outputs using AQUA. It includes:

- **Catalogues**: YAML-based mappings of models, experiments, and data sources for scalable and consistent data access.
- **Inference & Alignment Scripts**: Generate emulator outputs from model checkpoints (Anemoi) and align them with reference data.
- **Diagnostic Execution Logic**: Shell scripts and YAML configs to run AQUA diagnostics on selected emulator–reference pairs.
- **Supporting Resources**: Fixer rules and regridding weights for harmonised, comparable inputs across datasets.

5. Deployment on HPC Infrastructures

The evaluation framework supports **deployment across multiple HPC systems** and is fully containerised using **Singularity**, a solution well-suited for HPC environments. Singularity containers (provided directly with AQUA) ensure **compatibility with each AQUA release**, simplifying integration. This setup offers key advantages: consistent runtime environments, separation of inference and evaluation workflows, portability across systems, and compliance with strict HPC policies. By encapsulating dependencies and using **AQUA's official containers**, the **framework remains robust, reproducible, and easy to maintain**.

6. Extending Diagnostics for ML Emulation

To support more targeted evaluation of ML-based climate emulators, the framework is being **extended with new diagnostics specifically focused on ML performance.** These are developed following AQUA's modular structure, ensuring full compatibility and scalability. Initial additions include:

- **Root Mean Squared Error**: Measures spatial and temporal accuracy across variables.
- **Anomaly Correlation Coefficient**: Evaluates the emulator's ability to reproduce spatial patterns over time.

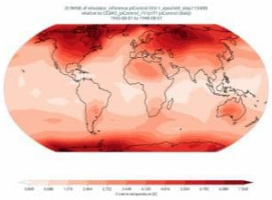
7. Performing the Evaluation

Configuration file for the RMSE diagnostic.

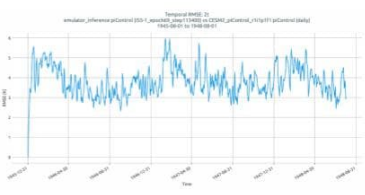
Configuration file for the ACC diagnostic.

In the current version, the evaluation suite includes the following diagnostics:

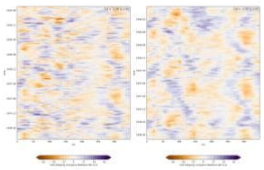
- Time series comparison
- RMSE
- ACC
- Global biases
- Teleconnection Diagnostics
  - NAO
  - ENSO
  - MJO



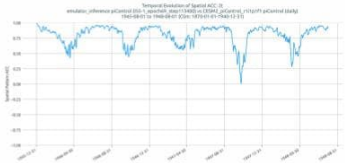
Spatial distribution of the RMSE



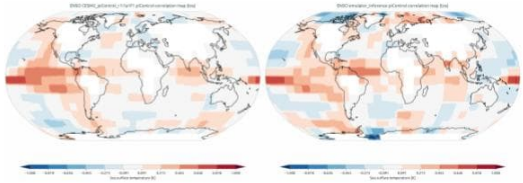
Temporal evolution of the RMSE



Hovmöller diagrams of outgoing longwave radiation proxies across the equatorial belt.



Temporal evolution of the ACC



Correlation map of the ENSO index with the SST

8. Future Steps

Ongoing work focuses on **scaling the framework to high-resolution data**, adding **new diagnostics** (e.g., SSIM, spectral metrics), and **contributing them back to AQUA**. Usability and accessibility will be enhanced through **curated documentation, example configurations, and CLI tools**, making the framework easier to adopt and extend by the community.

Funded by the European Union Destination Earth implemented by ECMWF esa EUMETSAT