Representing the 3D cloud radiative effects in ecRad using Machine Learning

Pélagie Alves(1), Christophe Bovalo(1), Matthew Chantry(2), Rémi Druilhe(1), David Guibert(1), Luca Marradi(1), Lionel Vincent(1)

(1) Eviden/Bull, France (2) ECMWF, UK

Contact : pelagie.alves@eviden.com

Highlights

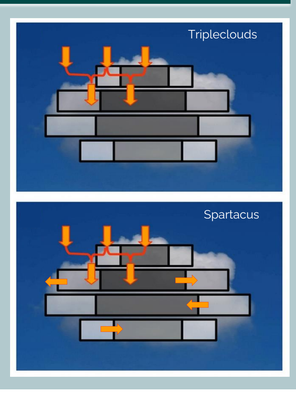
- **3D radiative cloud effects are well captured by deep learning models, especially the RNN**
- **All atmospheric columns are considered : with and without clouds (increasing the task complexity)**
- **Trainings were carried out on a heterogeneous cluster in multi-node, with Milan CPUs and Nvidia A100 GPUs**
- **Comparison of different model architectures (CNN-1D, Unet-1D and RNN)**
- **Two coupling strategies have been implemented**
- **Coupling strategies still need to be improved by adding distributed inference to enable scaling on GPUs**

1. Radiative schemes in ecRad

ecRad *(Hogan and Bozzo, 2016)* is a library used in IFS to compute vertical profiles of solar (**shortwave**) and near-infrared (**longwave**) fluxes and heating rates.
The radiative scheme is simulated through 5 different solvers. In this use case we will focus on **Tripleclouds** and **Spartacus.** Spartacus is an extension of Tripleclouds, representing the 3D cloud radiative effects but it is computationally more intensive.
The aim of this project is to **improve the accuracy of Tripleclouds** by training a neural network to emulate the 3D cloud radiative effects, based on the difference of the outputs between Tripleclouds and Spartacus.

$$SPARTACUS = Tripleclouds + \epsilon$$

where ϵ is the 3D cloud radiative effects and will be learnt by a neural network.
This development in ecRad is intended to be reintegrated into **RAPS20** (IFS-like model).
The present work pursues ***Meyer et al.***'s first attempt to correct Tripleclouds fluxes using SPARTACUS.



2. Bias correction with Deep Learning

2.1 Training strategies

Dataset

The dataset is generated by ECMWF for the **MAELSTROM** project. The **IFS model** is run every 30 days for 30 days and saving inputs/outputs every 25 hours. The forecasts were generated on a 40-km grid (TL511) resulting in 271,360 atmospheric columns every timestep.
Three datasets are available for training and validation the full 2020 year and 4 forecasts in 2019, respectively.
For our trainings, we used a subset of this dataset (28,221,440 columns for training, 203,520 for validation and 814,080 for testing).

Model Architecture

Model input data are the same as Tripleclouds.
The model returns the correction to be applied to the shortwave flux, the longwave flux, and their heating rates.

For this use-case we tested three model architectures (all with **attention mechanisms**):

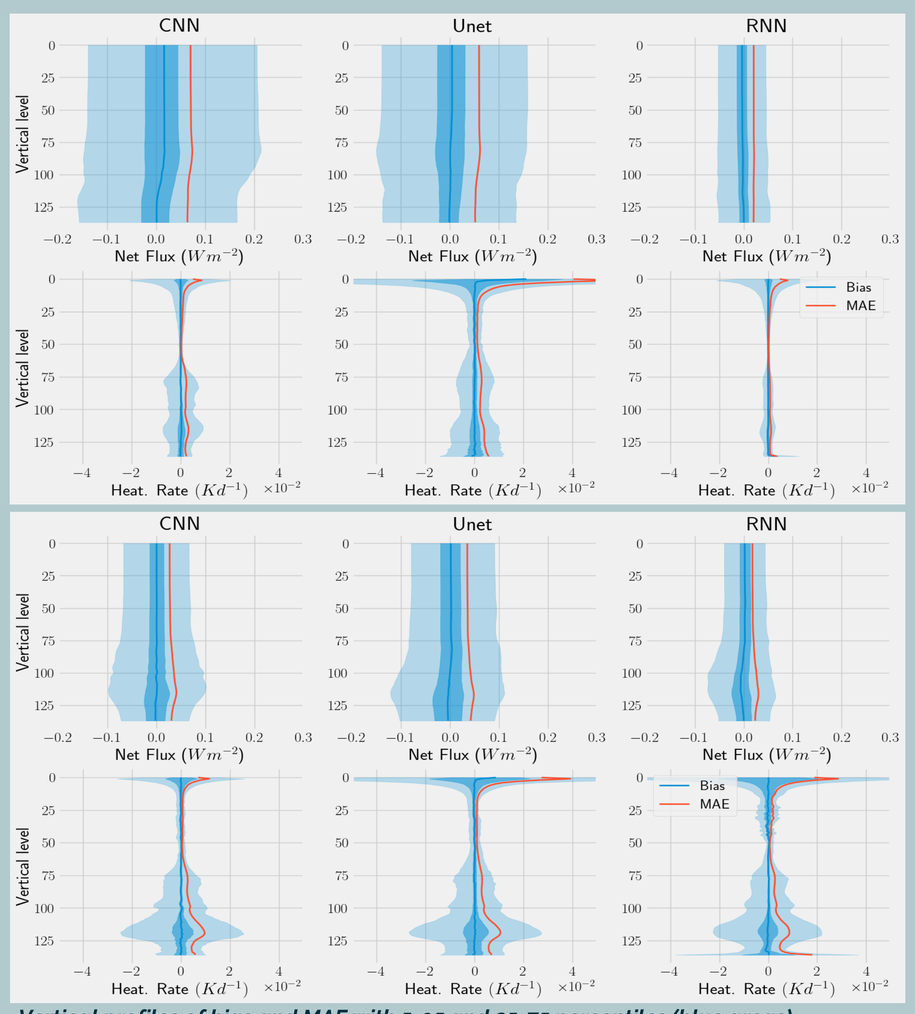
- **CNN-1D** using different dilation rates to propagate information
- **Unet-1D** with padding to keep the vertical structure
- **RNN (x2)** following *Ukkonen (2022)* and *Chantry (2022)*

Training

Trainings were performed using **TensorFlow** and **Horovod**.
Between 20 and 40 A100 GPUs were used per training, on a dataset of **~500GB**.
We chose to predict the corrective terms for the shortwave and longwave fluxes in addition to the heating rates using a custom layer (the heating rate is the vertical divergence of the net flux).
This is translated in a combined loss: $\mathcal{L} = \alpha\,\mathcal{L}_{flux} + \beta\,\mathcal{L}_{HR}$
where α and β are used to correct the imbalance between the $\mathcal{L}_{flux}$ and $\mathcal{L}_{HR}$.
Clear-sky and cloudy scenes are used in training so are the Earth dark side scenes for the shortwave flux.

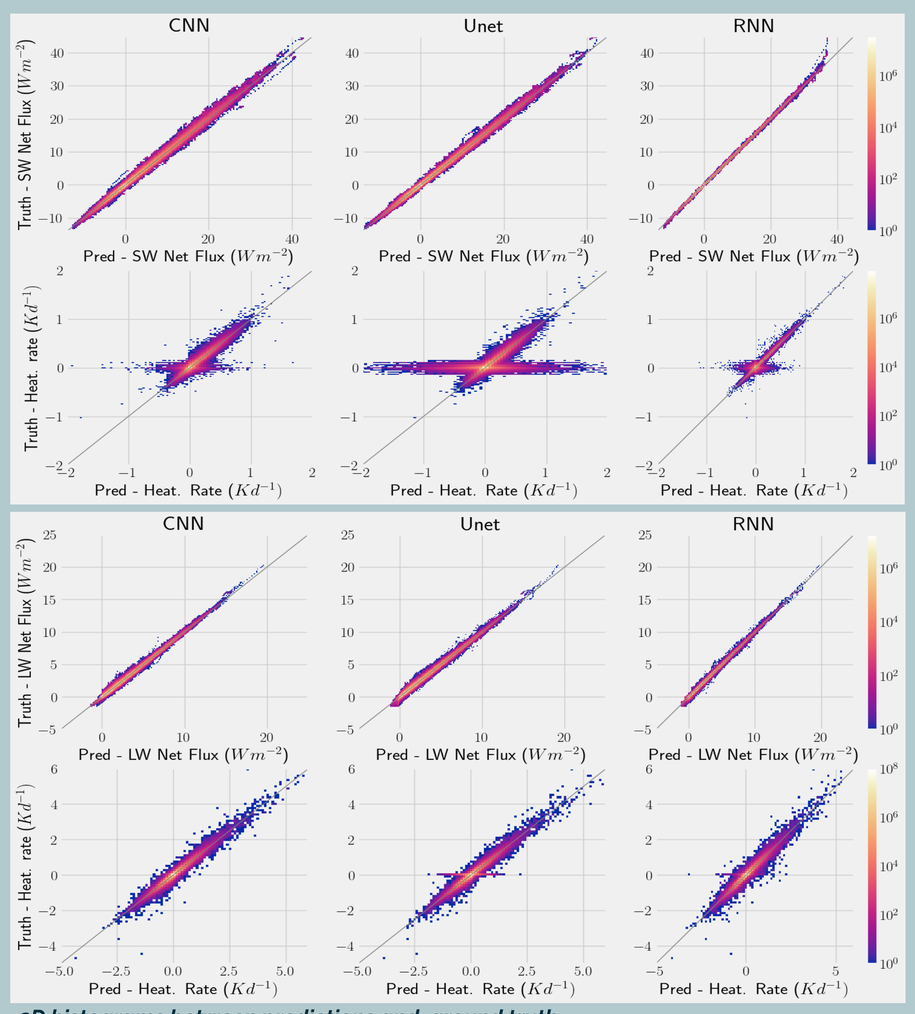
2.2 Results

Offline evaluation



Vertical profiles of bias and MAE with 5-95 and 25-75 percentiles (blue areas)

- RNN very good to correct both fluxes and heating rates
- CNN and Unet better to correct the longwave flux
- High errors at the top of the atmosphere for heating rates (pressure differences are small)



2D histograms between predictions and ground truth

- The error is well captured for flux by all the models, especially by the RNN
- Heating rates are more scattered
- High errors around zero for heating rates (pressure differences are small)

%	SW Net Flux	SW Heating Rate
CNN	8.23	51.26
Unet	6.81	67.36
RNN	2.43	18.50

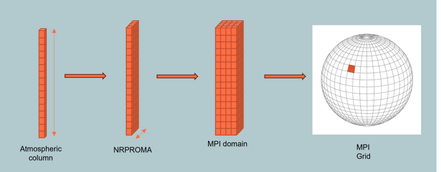
%	LW Net Flux	LW Heating Rate
CNN	3.61	18.44
U-Net	4.58	26.22
RNN	2.51	23.81

Mean Absolute Percentage Error

- Bulk errors are relatively small for fluxes
- MAPE < 7 % mean that the neural networks capture more than 93 % of the 3D radiative cloud effects
- RNN very good to correct both fluxes and heating rates
- CNN and Unet better to correct the longwave flux and especially heating rates

3. Coupling ecRad with ML models

To apply a correction to Tripleclouds considering the 3D cloud radiative effects, we need to establish communication between ecRad's *radiation_scheme* function written in Fortran and the AI inferer running in Python. This function is in a parallelized **OpenMP loop**. Two strategies were considered to solve this problem are **loose coupling** and **tight coupling**, both are still under development to improve scalability.



2.1 Loose coupling

This coupling is inspired by the client-server protocol, the MPI process in charge of the inference considered as the server and processes in charge of performing the solver calculation considered as clients. An effort has been made to transmit data using **MPI RMA**, enabling communication between the solver and the inferer.

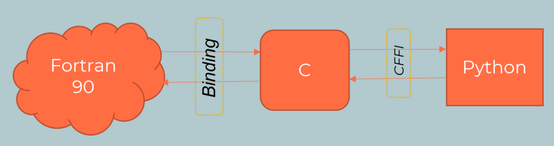
- Solver and inference are calculated asynchronously by different MPI processes
- Two synchronization points are required : one before starting inference, and another before adding the bias correction to Tripleclouds
- Works on heterogeneous architectures (CPU-GPU)

Improvements :

- Add the distributed inference on multi-GPU and multi-node
- Handle multi-threads with openMP
- Improve the pre/post-processing of the data between the inferer and the solver

2.2 Tight coupling

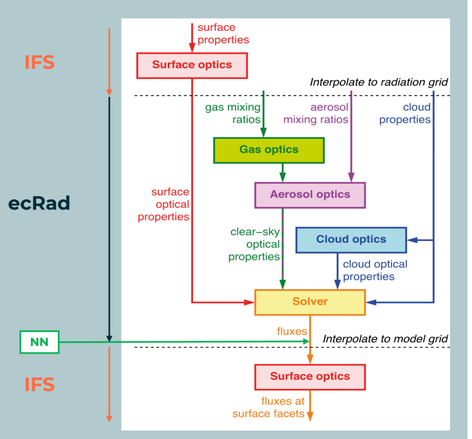
The inference is performed directly in the *radiation_scheme* function by each openMP thread. An interface has been designed to switch from Fortran to C++, and to **C++ to Python** and vice versa.



- No additional communication time
- The inferer batch size depends on NRPROMA, which represents several atmospheric columns determined by the optimal parameters of the openMP loop in ecRad.
- Only works on CPU for the moment
- Philosophy similar to ECMWF coupling library **Infero**

Improvements :

- Add a GPU version
- Reduce the loading time of the ML model
- Increase the batch size for the inferer
- Simplify coupling and distribute inference by using an API like NVIDIA's TorchFort which can directly load PyTorch models in Fortran codes



Adapted from Hogan and Bozzo.

Conclusions

The main objective of this study was to **learn the 3D cloud radiative effects** with Machine Learning. Our results extend the work of *Meyer et al. (2022)*.
Using more relevant datasets, we trained 3 different neural network architectures:

- **1D CNN** (custom network with dilation rates to propagate the information in the atmospheric column),
- **1D Unet** (widely used architecture) and
- **RNN** (to mimic the formulation of Tripleclouds).

Following *Ukkonen (2022)* and *Chantry (2022)* but for this correction task. **RNN were found to be excellent candidates to capture the 3D cloud radiative effects.** These results tend to confirm the dependance of ML model architectures to the physical processes they try to emulate.

We also implemented **two coupling strategies** to integrate the ML models into ecRad. **We manage to correct Tripleclouds in ecRad and even in RAPS**.
A first step has been done but both strategies need to be improved to achieve performance and scalability.

References

- Chantry, M. (2022). Machine Learning for Parameterised Physics. ECMWF Annual Seminar.
- Hogan, R. J. and A. Bozzo (2016). ECRAD: A new radiation scheme for the IFS. ECMWF Technical Memorandum number 787, 35pp: http://www.ecmwf.int/en/elibrary/16901-ecrad-new-radiation-scheme-ifs
- Meyer, D. et al. (2022). Machine learning emulation of 3D cloud radiative effects. *Journal of Advances in Modeling Earth Systems*, 14, e2021MS002550. https://doi.org/10.1029/2021MS002550
- Ukkonen, P. (2022). Exploring pathways to more accurate machine learning emulation of atmospheric radiative transfer. *Journal of Advances in Modeling Earth Systems*, 14, e2021MS002875. https://doi.org/10.1029/2021MS002875