AI applications to Earth Science at the Barcelona Supercomputing Center

Carlos Alberto Gómez Gonzalez

ESA Phi-week 2021

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

STARS POST-DOCTORAL PROGRAMME

Who we are

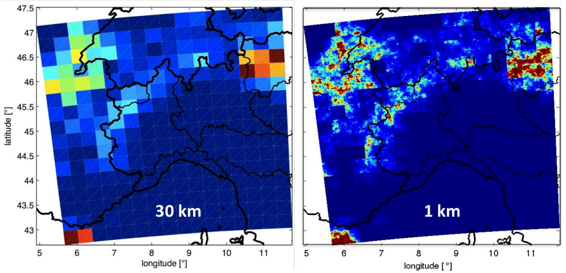
Barcelona Supercomputing Center (BSC)

Me: Postdoctoral fellow at the Earth Sciences department of the BSC working on Artificial Intelligence (AI) and Machine Learning applications

- BSC is Spain’s leading supercomputing and HPC center
- It hosts the MareNostrum (one of the largest supercomputers in Europe) and other clusters, such as the Power-CTE (GPU-based computing)
- ES department hosts 100+ researchers with expertise in climate science, atmospheric composition, and software and data engineering

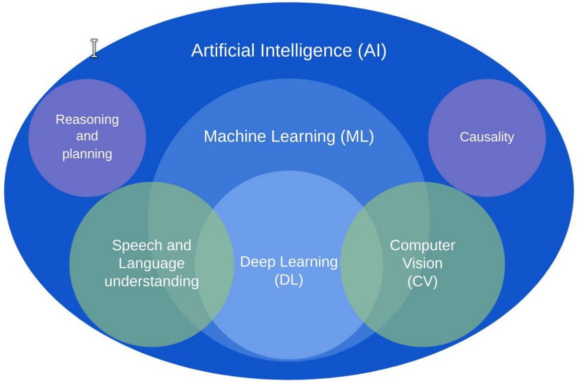
Cross-disciplinarity

Climate Science



- Problem definition
- Domain expertise
- Baseline approaches
- Data sources identification
- Validation metrics

Artificial Intelligence



- Framing Earth science problems from AI perspective
- Identification and development of AI approaches for ES needs

AI and ML engineering

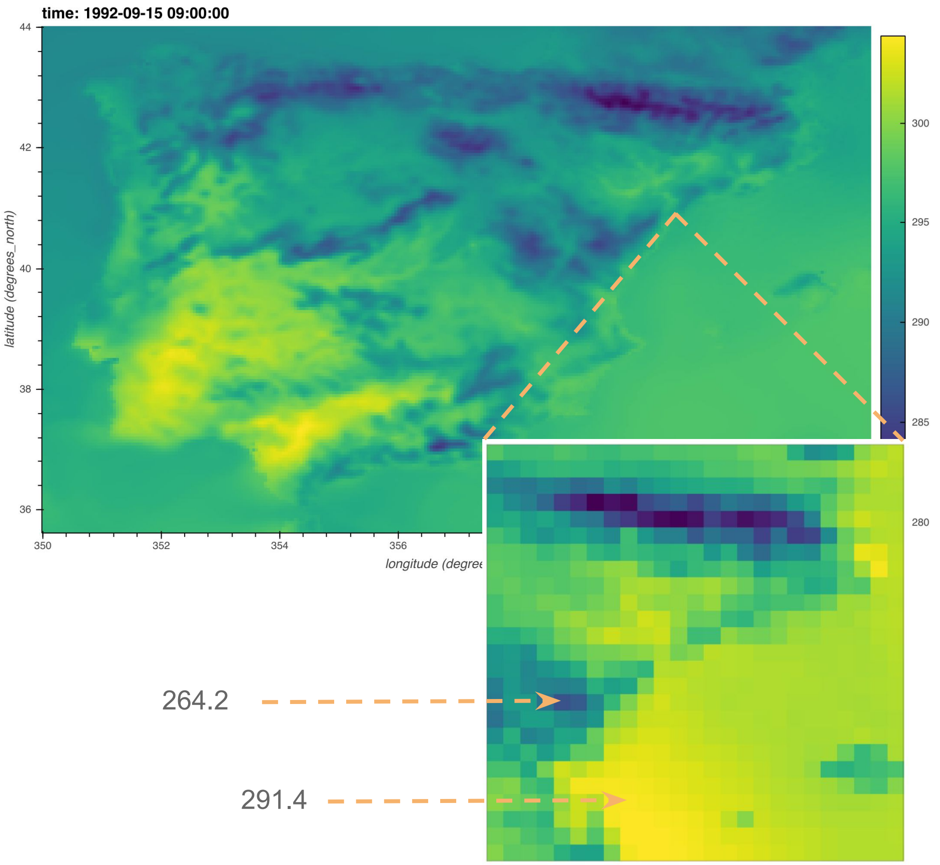
```
import numpy as np
import tensorflow as tf
from tensorflow.keras.layers import (Add, Conv2D, Input, Dropout, GaussianDropout,
                                     Concatenate, UpSampling2D, TimeDistributed,
                                     LocallyConnected2D, Conv2DTranspose)
from tensorflow.keras.models import Model

from .blocks import Conv3DBlock

```

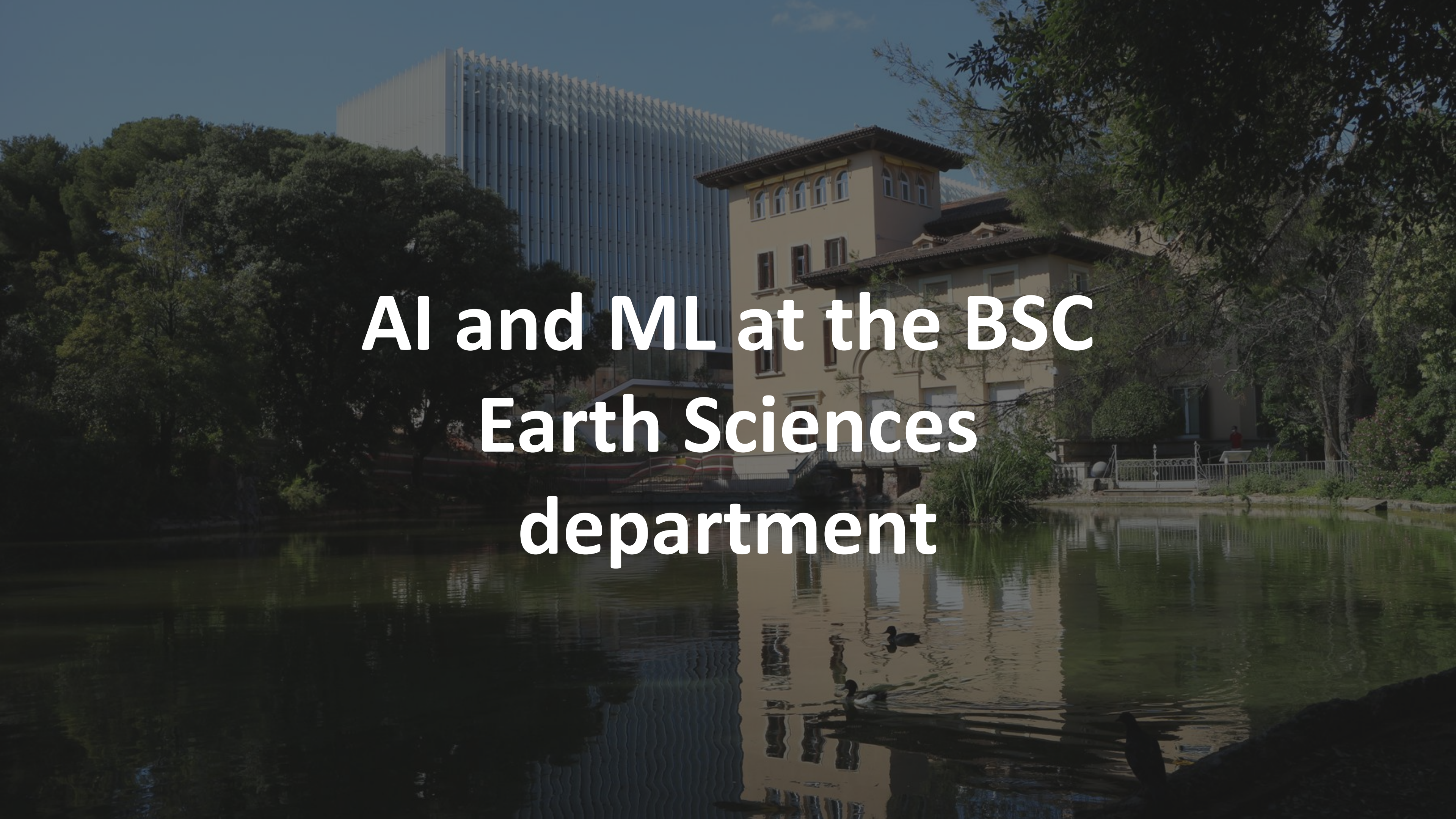
- Development of robust, efficient and open code
- Smart testing and model design/tuning
- Reusability and reproducibility

AI for Earth Sciences

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

- Spatio-temporal processes
- Common structural prior (grids and time series)
- Common tasks (AI <---> ES):
  - Time series forecasting → regression
  - Next frame video prediction → multidim. regression and forecasting
  - Super-resolution → stat. downscaling
  - Object recognition → pattern finding
  - Inpainting → missing data filling
  - Image to image translation → transfer functions, regression
- Open challenges: Uncertainty, causality, explainability





AI and ML at the BSC Earth Sciences department

ML for atmospheric composition

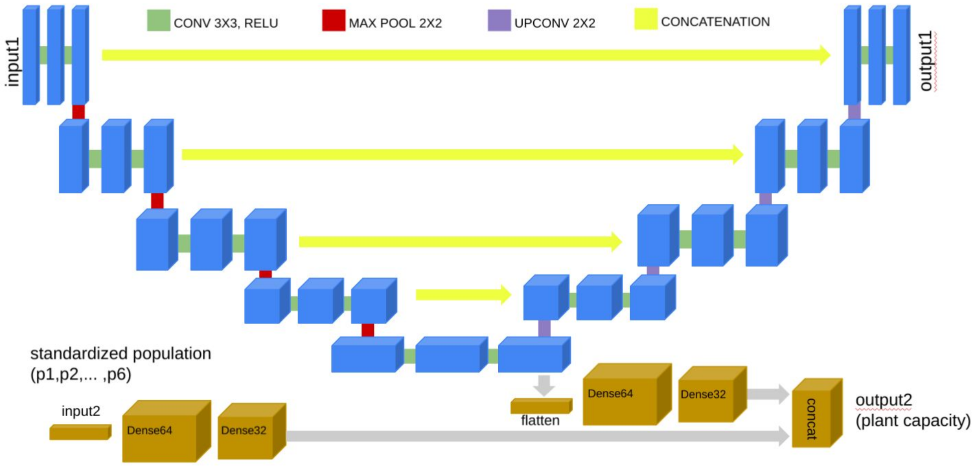
- Correction of the CAMS O3 (ground level ozone) ensemble forecasts over Spain
    - Comparison of different statistical bias-correction methods with ML-based approaches, such as gradient boosting machines
- Machine Learning-based estimation of surface NO2 concentrations from satellite data
- Bias correction of global CAMS reanalysis dataset of atmospheric composition using surface Air Quality observations

AQ-WATCH

EARLY-ADAPT

Computer vision for improving emission models

- Wastewater treatment plants (WWTPs) are a source responsible for ~20% of the CH_4 emissions
- Developing a methodology for the localization and characterization of WWTPs using a UNET-based algorithm
- The model is trained on Sentinel-2 10m images over the Iberian peninsula
- This methodology is useful for improving the spatial proxies for distributing CH_4 emissions in atmospheric emission models

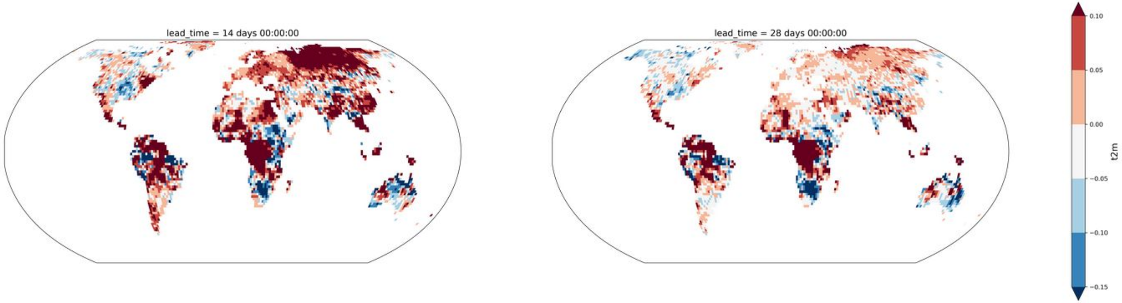


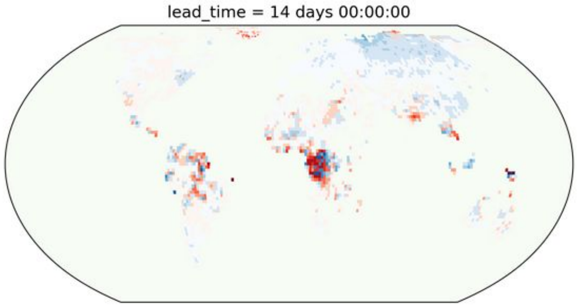
Sentinel-2 | Prediction | Groundtruth

Sub-seasonal to seasonal AI challenge

- Work in the context of the S2S-AI data challenge
- Improve subseasonal-to-seasonal precipitation and temperature forecasts with Machine Learning/Artificial Intelligence
- Objective: to improve week 3-4 and 5-6 subseasonal global probabilistic 2m temperature and total precipitation tercile forecasts (2020)
- Different ML-based classifiers are applied: logistic regression, random forest and gradient boosting algorithm, as well as a multi-approach ensemble

RPSS Random Forest





AI for the improvement of seasonal forecasts

Empirical model for biomass burning emissions

CONFESS

Consistent representation of temporal variations of boundary forcings in reanalyses and seasonal forecasts

https://confess-h2020.eu

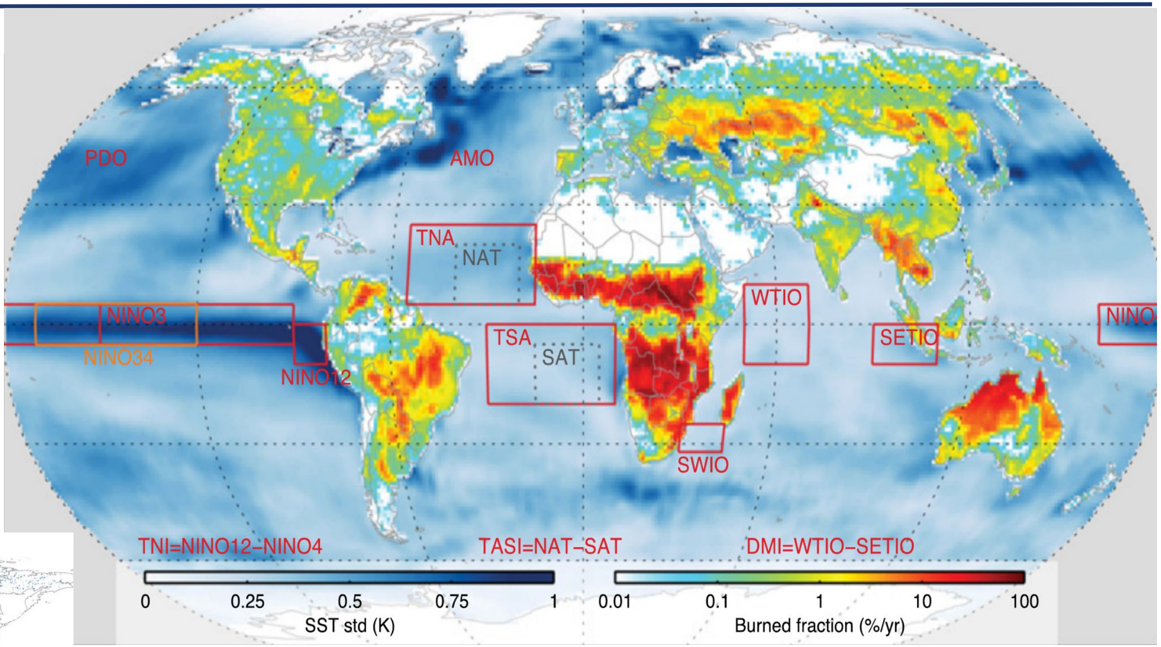
Objectives of this sub-task:

- to develop and validate a Machine Learning model for monthly burned area using as predictors:
    - large-scale indices
    - biomass and land cover data
- to apply the model developed for the BA to model the monthly total emissions per species (GFAS)

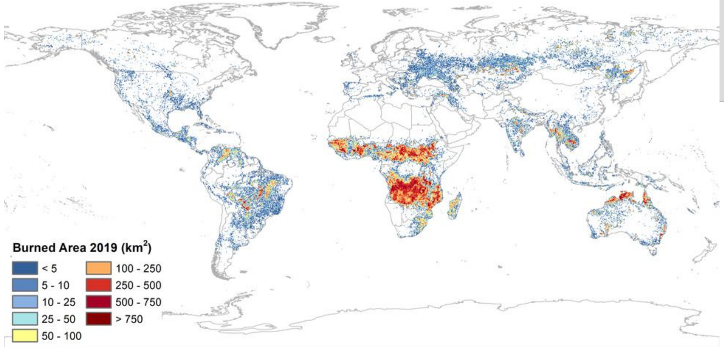
Data for burned area model (I)

Subset of the Ocean Climate Index (OCI) regions used

Annual burned fraction (1997-2014) is shown on land



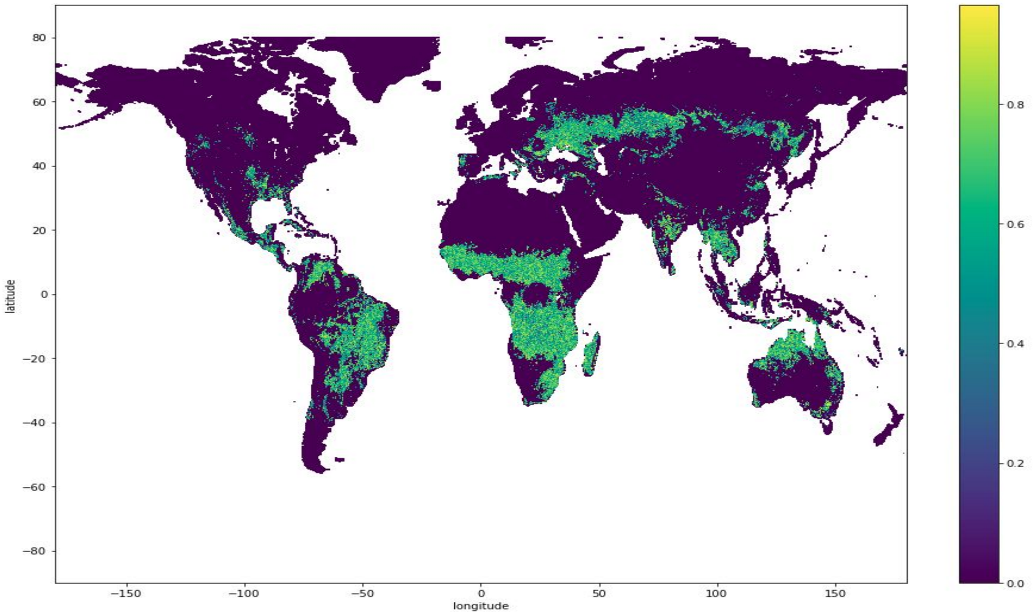
Taken from Chen et al., 2016



ESA Fire CCI 5.1 burned area data regridded to 0.25º

Data for burned area model (II)

- 0.25 deg resolution burned area aggregated over 12 months periods centered around peak month
- Grid cell will more than 50% no fire years were removed
- OCIs as predictors with varying lags (months to the past)
- Empirical correlations (> 0.6) between OCIs and burned area (figure on the right)



Modelling the burned area

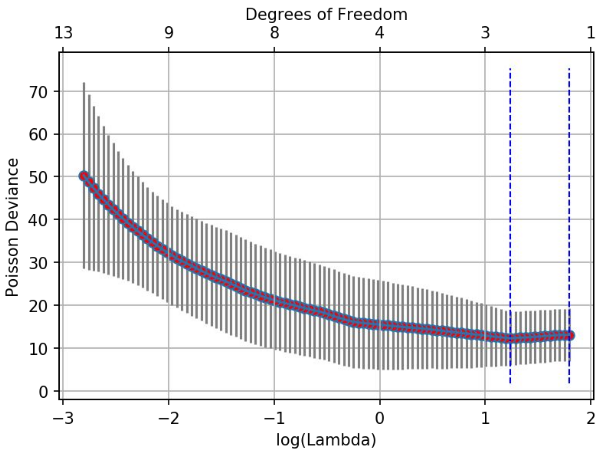
Barcelona Supercomputing Center
Centro Nacional de Supercomputación

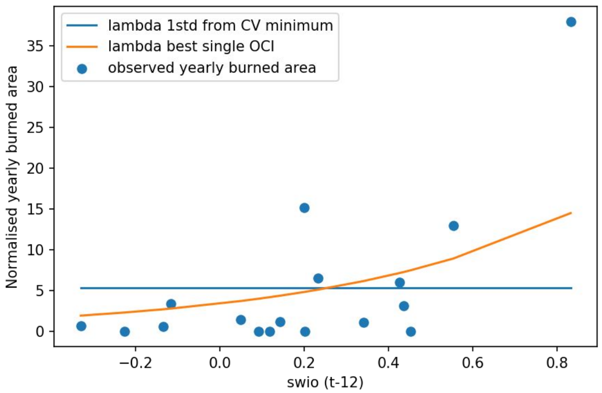
- We model the burned area using poisson regression where the mean is assumed to be a linear function of OCI (x) with a log link transform

$$\log \mu(x) = \beta_0 + \beta' x$$

- Beta parameters are obtained by minimising poisson likelihood function together with a sparsity inducing Lasso penalty

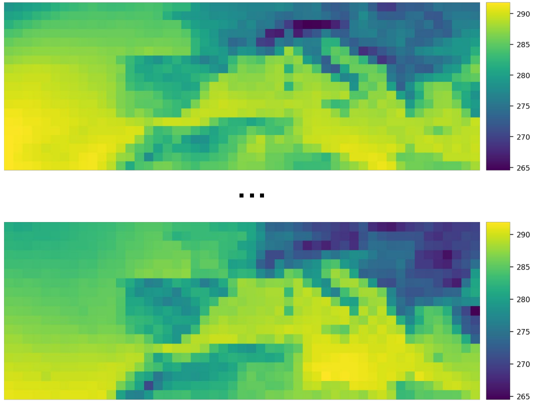
$$\min_{\beta_0,\beta} -\frac{1}{N} l(\beta|X,Y) + \lambda \sum_{i=1}^{N} |\beta_i|$$



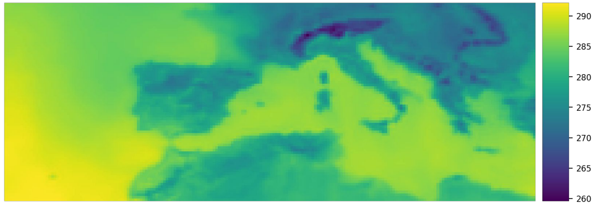


Post-processing dynamical seasonal forecasts

- Dynamical seasonal prediction systems have limited capability in reproducing the precipitation and temperature at seasonal to annual lead times
- Their spatial resolution is too coarse for many applications
- Resolution in EO depends on the satellite orbit configuration and sensor design while for ES dynamical models is a matter of computational budget



Dynamical forecast ensemble members



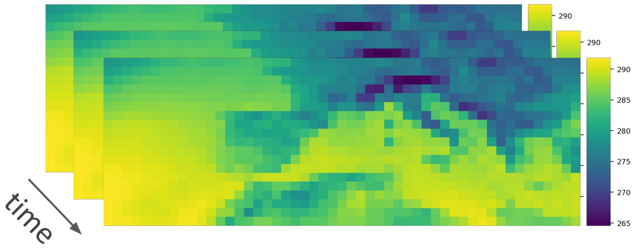
Bias corrected and downscaled seasonal grids

Climate gridded data

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

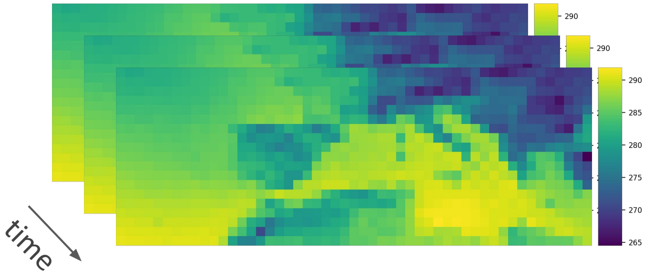
Daily surface air temperature grids from 1981 to 2020 (~15k time samples) for the Mediterranean region

Seasonal hindcast (SEAS5 tas 1º)



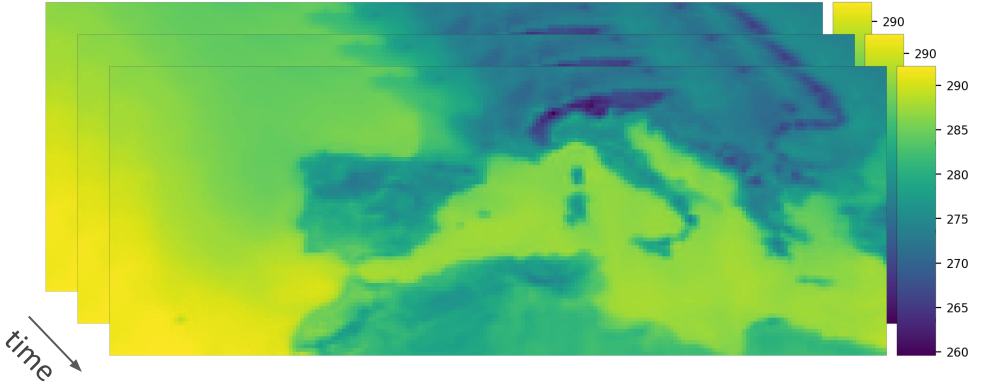
Ensemble member 1

...

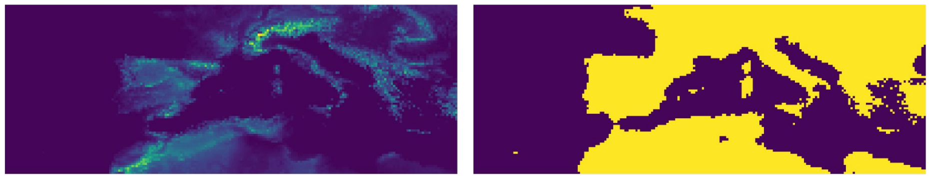


Ensemble member n

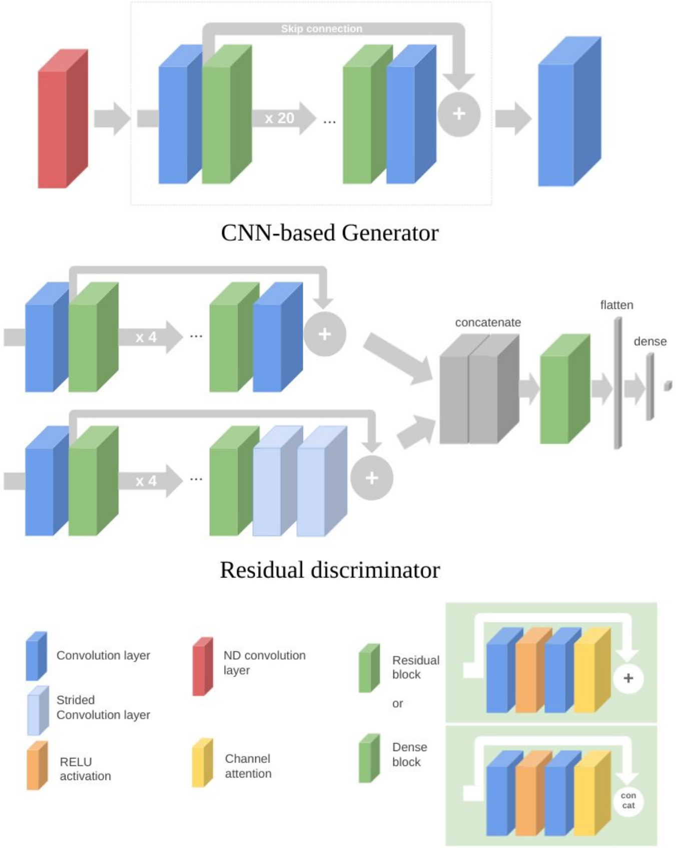
Observational reference (ERA5 tas 0.25º)



Time independent fields (topography)

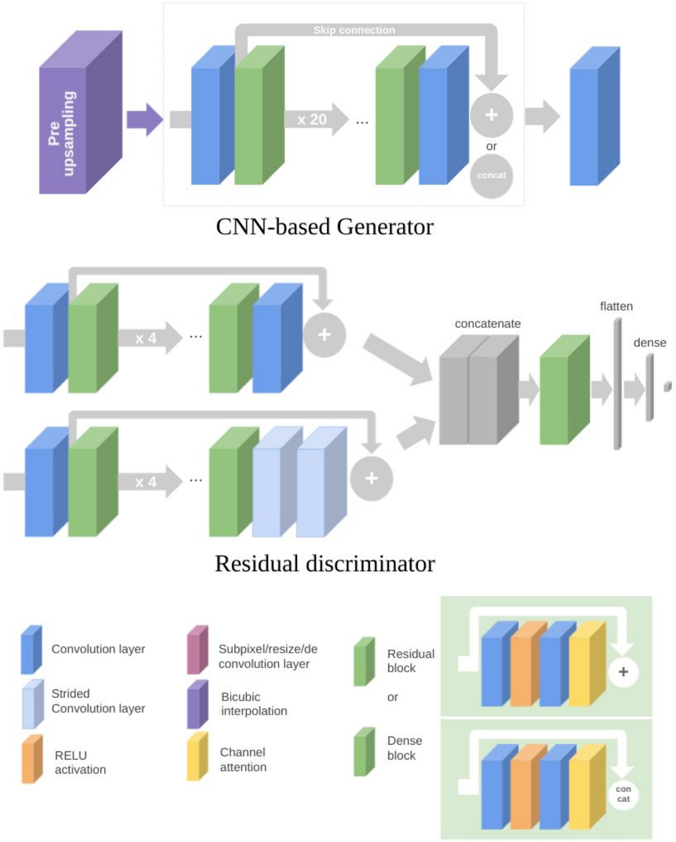


DL-based model for bias correction



- Simplified architecture of the CNN
- The network on the top can be trained alone in a fully supervised manner
- Both networks, generator and discriminator can be trained in a conditional adversarial fashion
- Models are trained on seasonal and reanalysis (observational) data
- A Python package based on Tensorflow is being developed to take advantage of HPC clusters for distributed GPU/CPU training

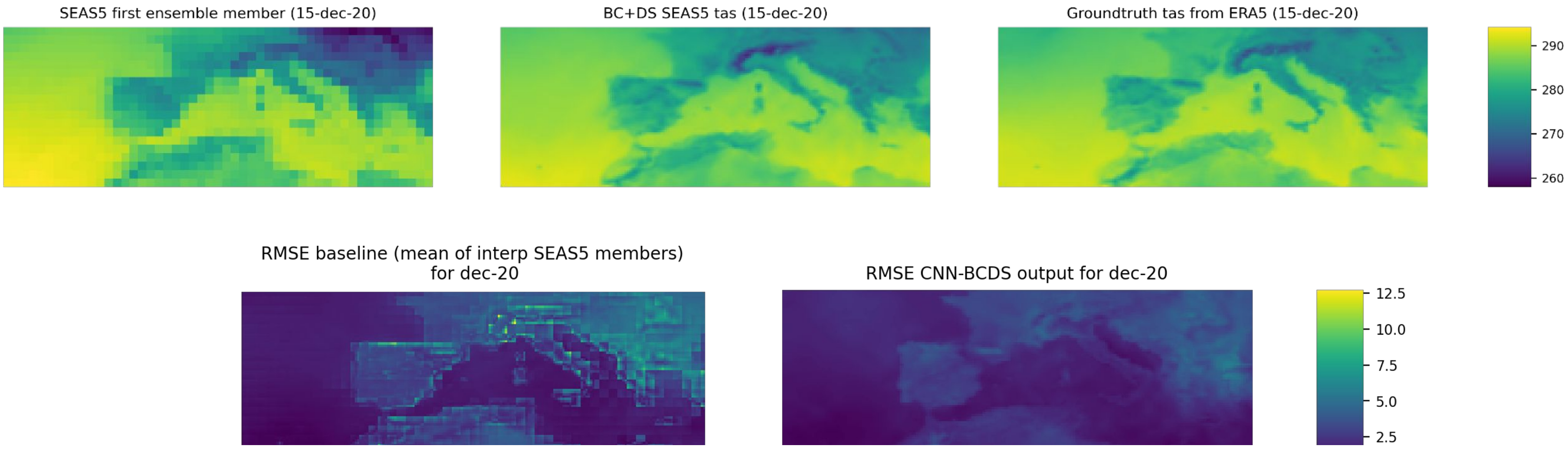
DL-based model for downscaling



- Trained on reanalysis/observational data
- Training pairs are created by downsampling and upsampling the high-resolution reanalysis data to create a HR and LR training pairs
- The model incorporates auxiliary information, such as topography
- A Python package based on Tensorflow is being developed to take advantage of HPC clusters for distributed GPU/CPU training
- This algorithms can be applied to downscale/super-resolve any gridded climate/EO dataset

Post-processed seasonal data

- Preliminary results show the post-processed seasonal temperature grids with improved spatial resolution and lower RMSE wrt the raw SEAS5



Thanks!

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

- Synergies between Earth Science, AI/ML and HPC are far from easy but possible
- A few examples of such synergies at the BSC-ES were showcased
- The AI for ES field is going through a prolific proof-of-concept phase
- Next steps:
    - more emphasis on data and software engineering for the creation of tools to foster reproducibility and replicability
    - AI for process understanding (causality, explainability)
    - physics-driven AI

Acknowledgements:

The research leading to these results has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement H2020-MSCA-COFUND-2016-754433 and the EU H2020 Framework Programme under grant agreements n° GA 823988 (ESiWACE-2), GA 869575 (FOCUS-Africa), GA 869565 (VitiGEOSS) and GA 101004156 (CONFESS)..

carlos.gomez@bsc.es