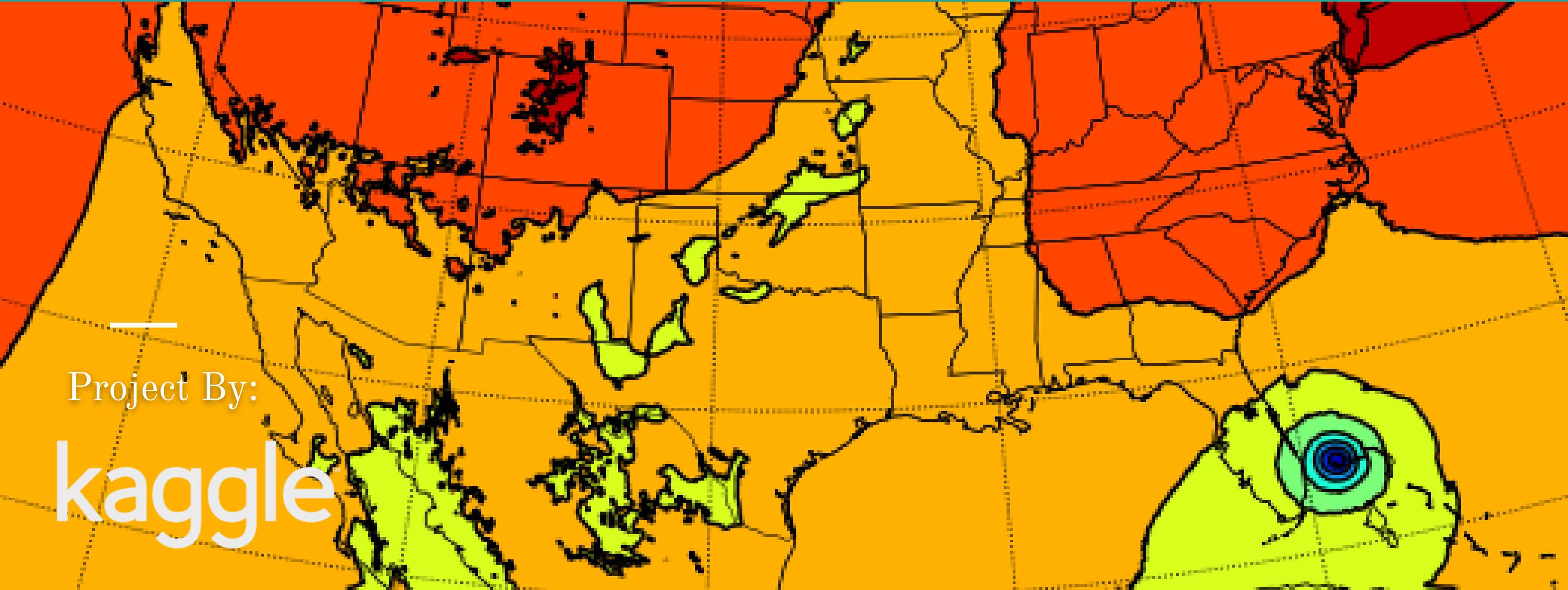
Solar Energy Prediction

Iason Tsardanidis



Project By:

kaggle

Project outline:

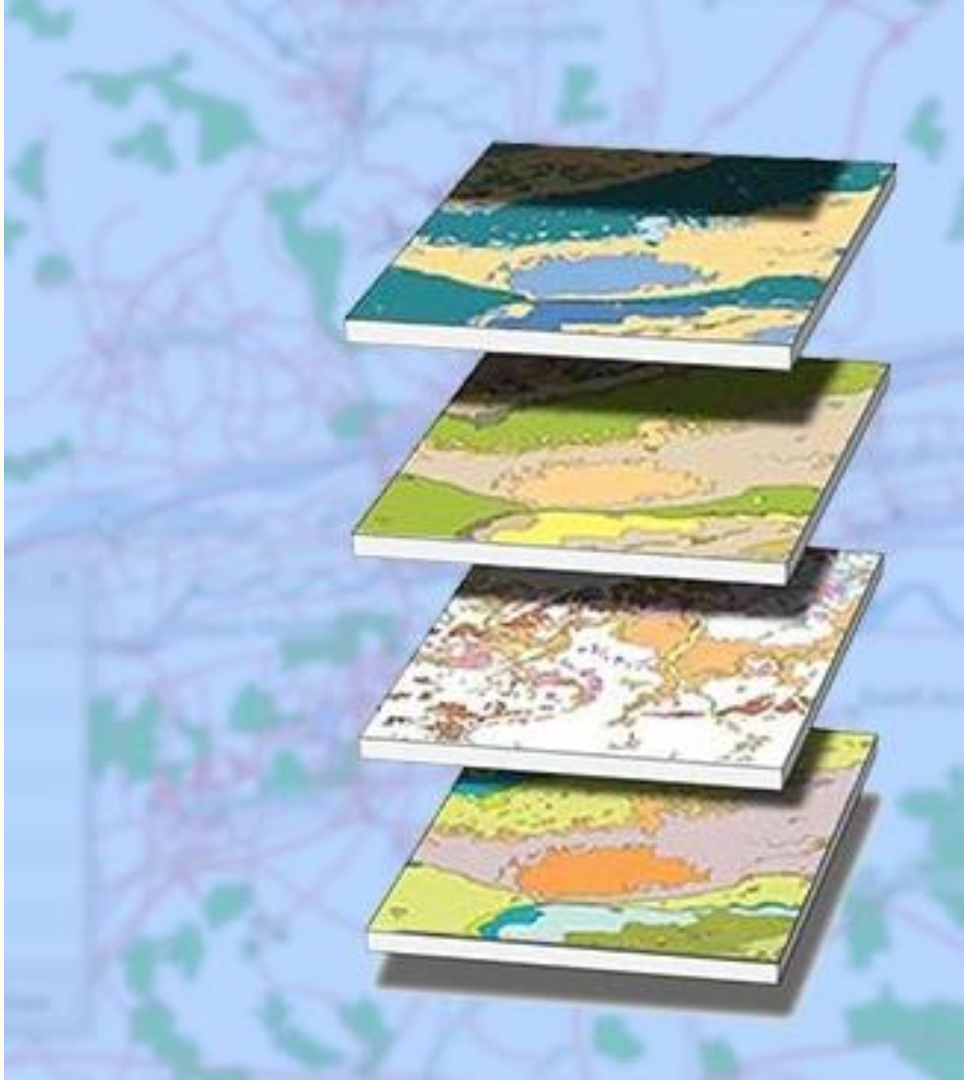
- ❑ Data Description
- ❑ Methodology Approached
- ❑ Exported Model
- ❑ Results Evaluation
- ❑ Possible extension

Data Description

NetCDF DATA

(network Common Data Form)

Data format for storing multidimensional data. This data can be temperature, humidity, pressure, wind speed and direction in both vector and raster format. Each variables can displayed through a dimension in GIS (Data can be for Atmospheric, oceanographic and earth sciences purpose) with layers or table from the netCDF file. The dataset comprises space (latitude ,longitude and altitude), time, etc.



Some Acknowledgments

Data provided by the American Meteorological Society Committees on Artificial Intelligence Applications to Environmental Science, Probability and Statistics, and Earth and Energy and the Input numerical weather prediction data for the contest comes from the NOAA/ESRL Global Ensemble Forecast System (GEFS).

Meteorological Data finds numerous of applications nowadays in Big Data statistics and accurate Meteorological Prediction consists a tough challenge since they form a very delicate complex system.

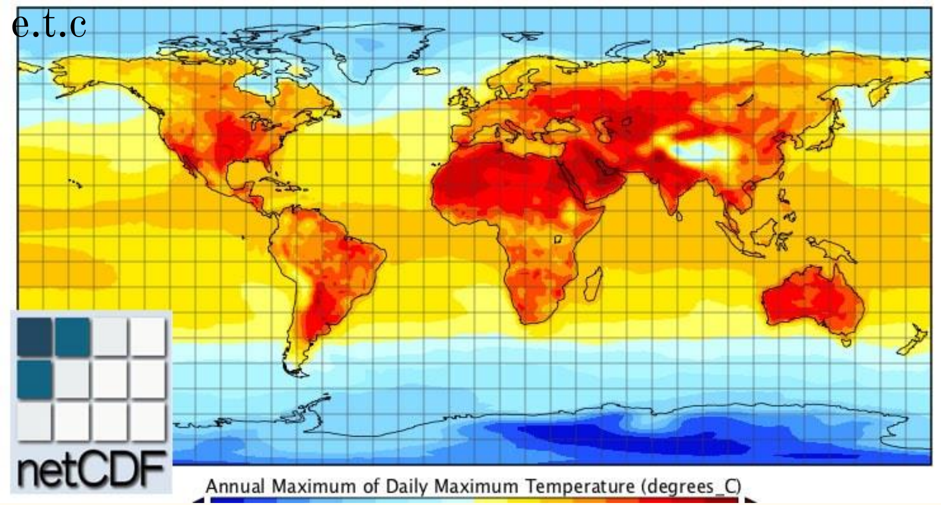
Our Dataset

- ★ The first dimension is the date of the model run and will correspond directly with a row.
- ★ The second dimension is the ensemble member that the forecast comes from.The GEFS has 11 ensemble members with perturbed initial conditions.
- ★ The third dimension is the forecast hour, which runs from 12 to 24 hours in 3 hour increments. All model runs start at 00 UTC, so they will always correspond to the same universal time although local solar time will vary over each year.
- ★ The fourth and fifth dimensions are the latitude and longitude uniform spatial grid.

Source: (https://www.igismap.com/multidimensional-data/)

We have 5113 samples for a period of 14 years from January of 1994 until December of 2007.

15 in Total NetCDF files for each meteorological
variable : Air Pressure, Total cloud cover, Max-Min Temperature, Upward-Downward short-long wave radiation, Precipitation Accumulated e.t.c



To sum up...

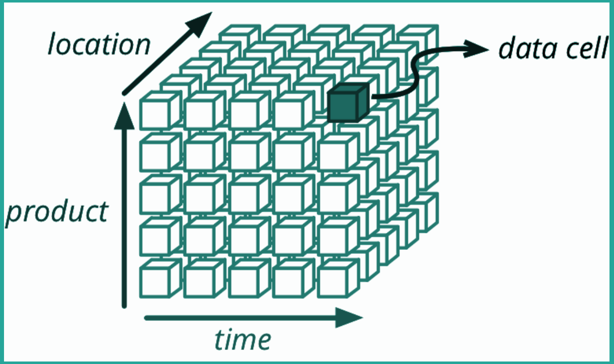
Inputs ....(It’s a mess):

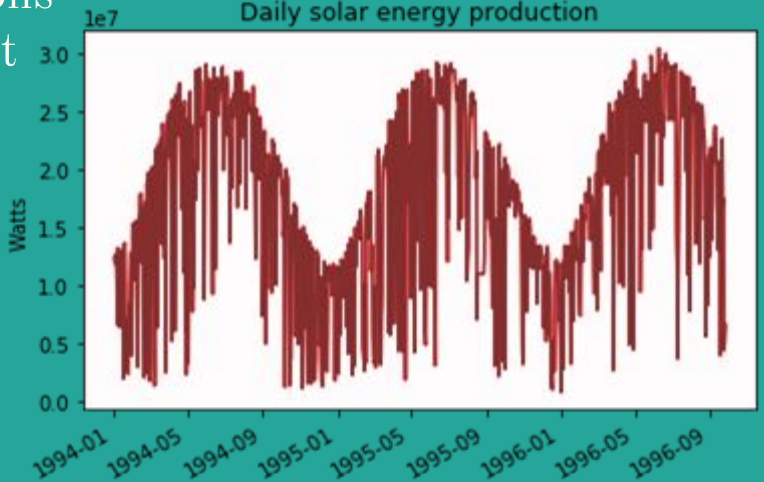
- 15 NetCDF files for each meteorological variable
- 5113 samples for a period of 14 years
- 11 ensemble methods for every GEFS model
- 5 individual forecasts for every 3 hours of every date
  - Sunlight hours
- 16 x 9 latitude vs longitude space model predictions
  - (small grid resolution) → We will use all of it

Outputs:

- The Total Solar Energy Produced per day.

High-Multidimensional Objects!!!



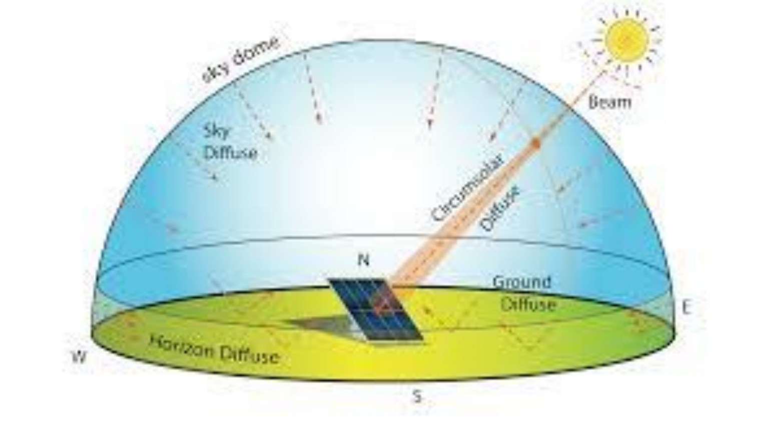


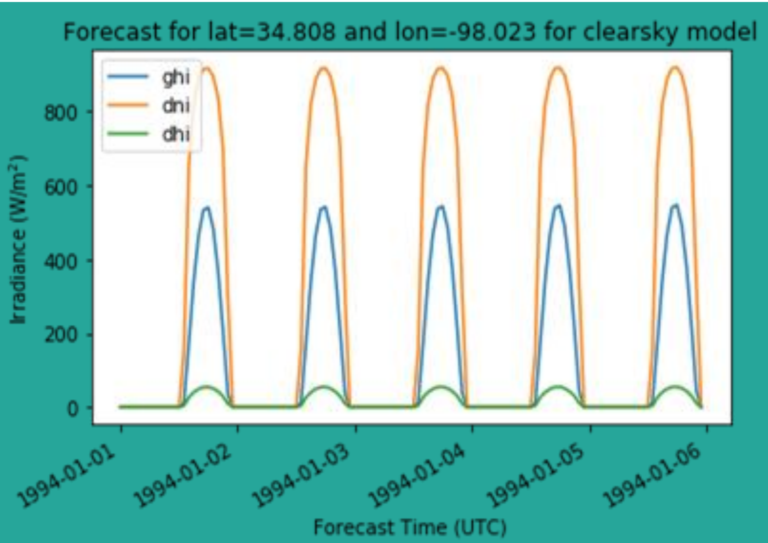
Data Engineering
(New Features Construction)

- Direct Normal Irradiance (DNI) is the amount of solar radiation received per unit area by a surface that is always held perpendicular (or normal) to the rays that come in a straight line from the direction of the sun at its current position in the sky.
- Diffuse Horizontal Irradiance (DHI) is the amount of radiation received per unit area by a surface (not subject to any shade or shadow) that does not arrive on a direct path from the sun, but has been scattered by molecules and particles in the atmosphere and comes equally from all directions
- Global Horizontal Irradiance (GHI) is the total amount of shortwave radiation received from above by a surface horizontal to the ground.

$$\mathbf{GHI = DNI \times \cos(\theta) + DHI}$$

(where θ is the solar zenith angle)





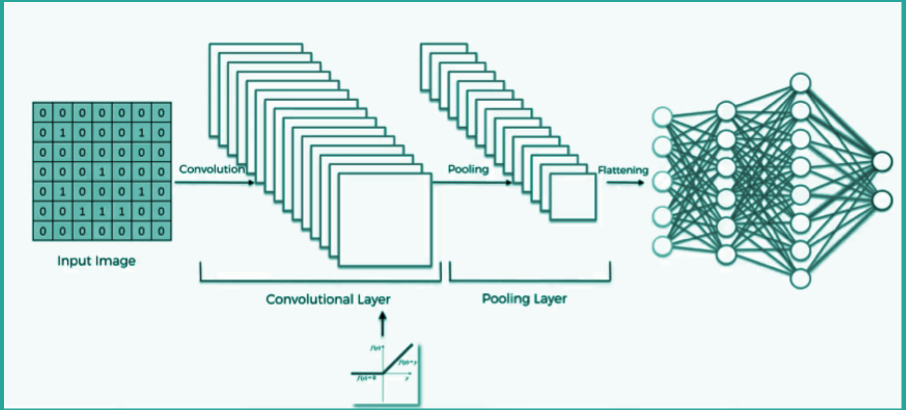
Methodology Approached & Exported Model

Convolutional Neural Networks (CNN)

- Class of deep neural networks, most commonly applied to analyzing visual imagery.
- Ideal for analyzing multidimensional input objects
- They let us keep spatial information - do not change the type of data
- Able to distinguish local features
- Transition-invariant
- We are able to reduce dimensionality (Pooling Layers)

Some disadvantages:

- High computational cost.
- If you don't have a good GPU they are quite slow to train (for complex tasks).
- They use to need a lot of training data.



0	0	0	0	0	0	0
0	1	0	0	0	1	0
0	0	0	0	0	0	0
0	0	0	1	0	0	0
0	1	0	0	0	1	0
0	0	1	1	1	0	0
0	0	0	0	0	0	0

Input Scaling - Normalizing

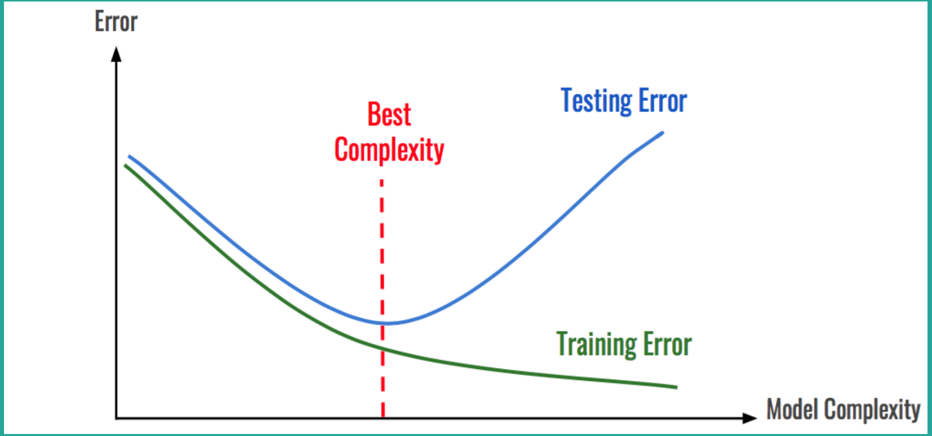
- Since we are dealing with NN → No Feature Selection necessary
- All values will be normalized in [0,1]
    - Target values will be divided with the park max daily production ~ 40 MW
    - Input values will be scaled via MinMaxScaler model (15 different scaler model for every variable)

Train - Test Sample Split

- Data until the December of 2004 will be used as training sample.
- Test sample → data from January of 2005 and after .

Overfitting Avoidance

- Grid-Search hyperparameters tuning
- Gaussian Noise Layer
    - Add (some) noise between layers
- Early Stopping
- Dropout (possible alternative)
- Reguralization(another alternative)



Source: stats.stackexchange.com

Ensemble Stacked 2D models Vs Stacked 3D models

2D Model

- Less complex, we just average 11 simpler models
- Less adapted to data (less danger of overfitting)
- Better Results for our task

- More time needed to train

3D Model

- Much quicker, we just train a whole very complex network
- Model Integrity → we don’t average

- Much More complex
- More trainable parameters
- Peril of overfitting

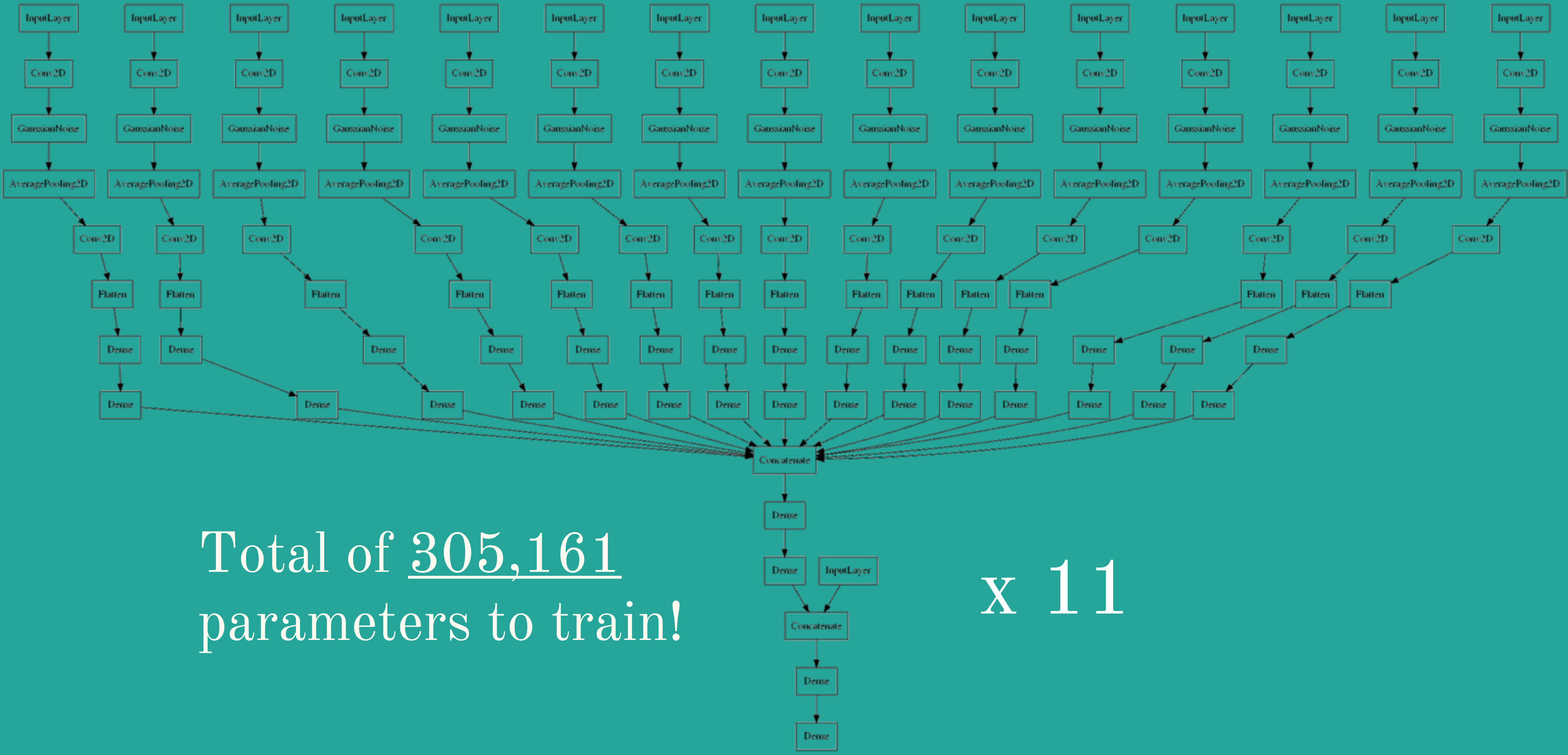
Model Architecture

(2D Ensemble Stacked CNN model)

- 11 individual stacked CNN models for each GEFS member
  - We <u>average</u> them at the end!!!
- Each Stacked CNN model composed by 15 2D CNN simpler models for every meteorological variable. *
  - After concatenation → there is a second layer of inputs for the extra features of (clearsky) radiation flux and the month of the year.
- Each 2D CNN has:
  - 2 conv2d layers
  - 1 gaussian noise layer
  - 1 Average Pooling Layer

* * Illustration of a stacked CNN model in the next page

(Model Illustration)



Total of 305,161 parameters to train!

x 11

Training Routines

After a (small) Grid Search Tuning:

Activation Function: **RELU** (except output layer → Linear)

Optimization Algorithm: **Adam**

Loss_Function: **Mean Absolute Error**

Number of Epochs: **25**

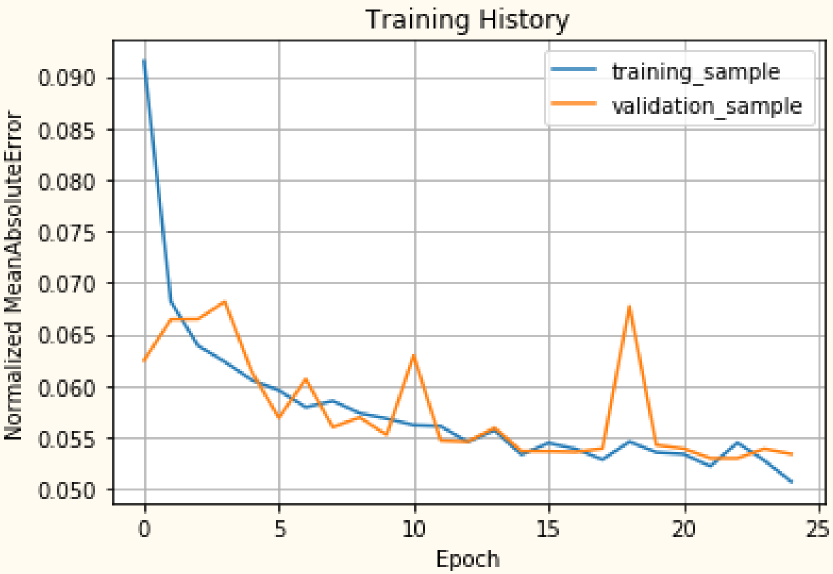
Batch_size: **25**

Kernel_Initialization: **Xavier Method**

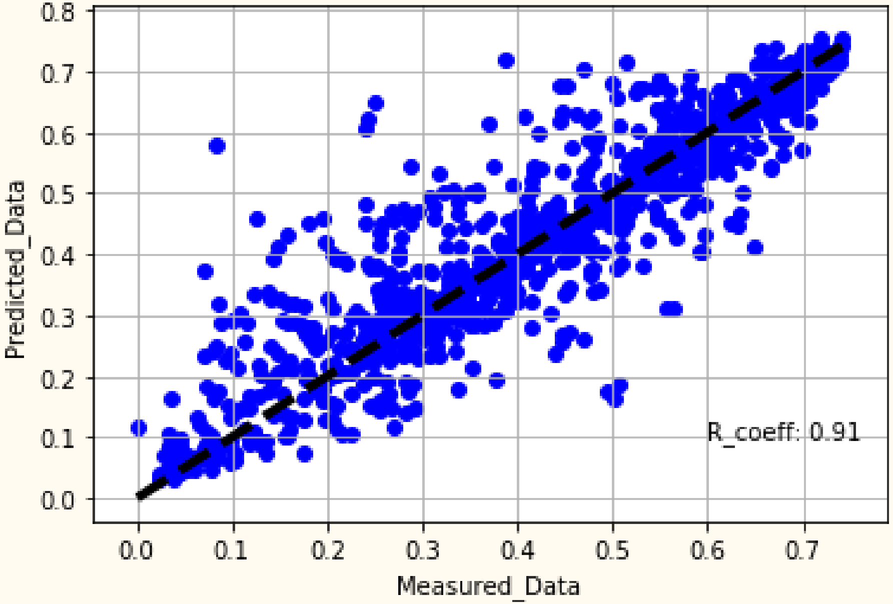
Results Evaluation

During Training

History of Estimator:

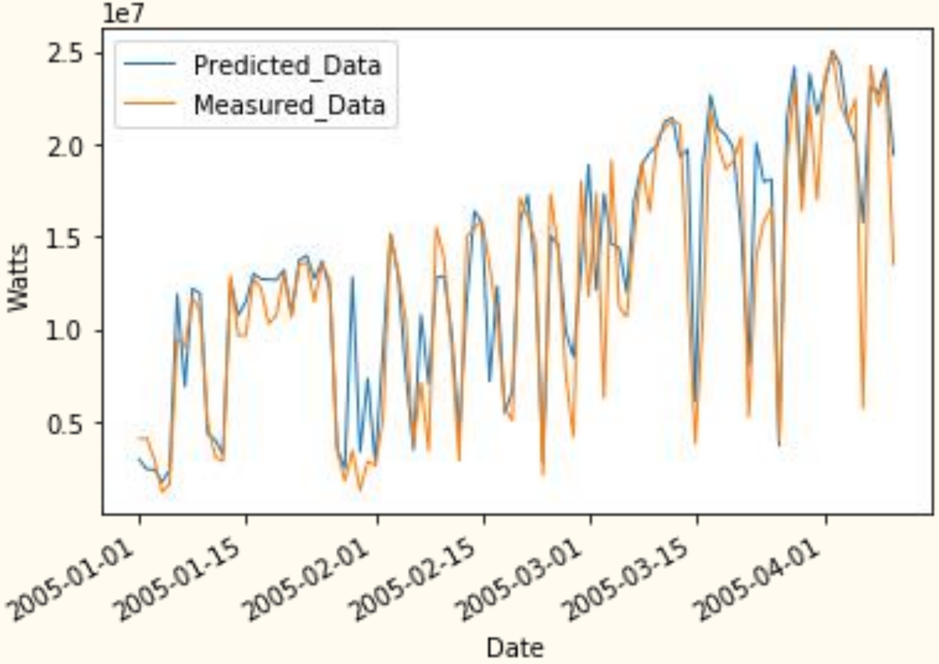


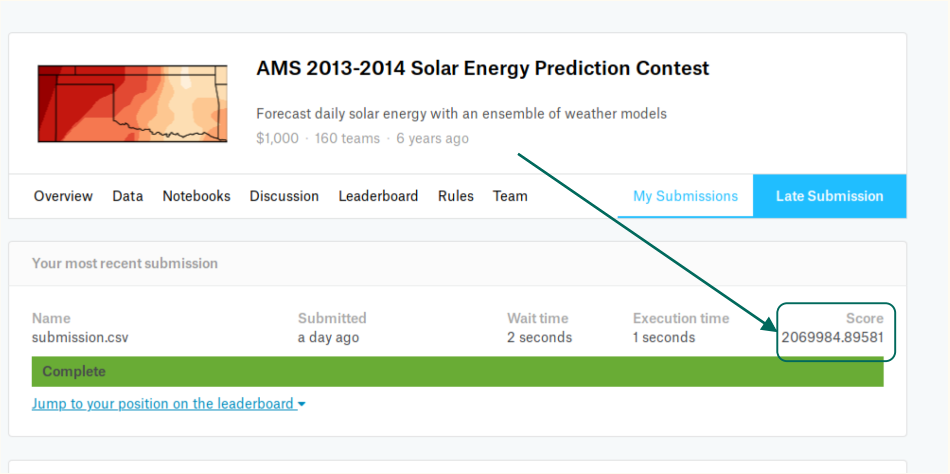
Correlation Plot:



MAE: ∼ 5% of Maximum Power

Evaluation of Results: Mean Absolute Error





AMS 2013-2014 Solar Energy Prediction Contest

Forecast daily solar energy with an ensemble of weather models

$1,000 · 160 teams · 6 years ago

Your most recent submission

Name	Submitted	Wait time	Execution time	Score
submission.csv	a day ago	2 seconds	1 seconds	2069984.89581

Complete

Jump to your position on the leaderboard

Computation Time: ~10 minutes

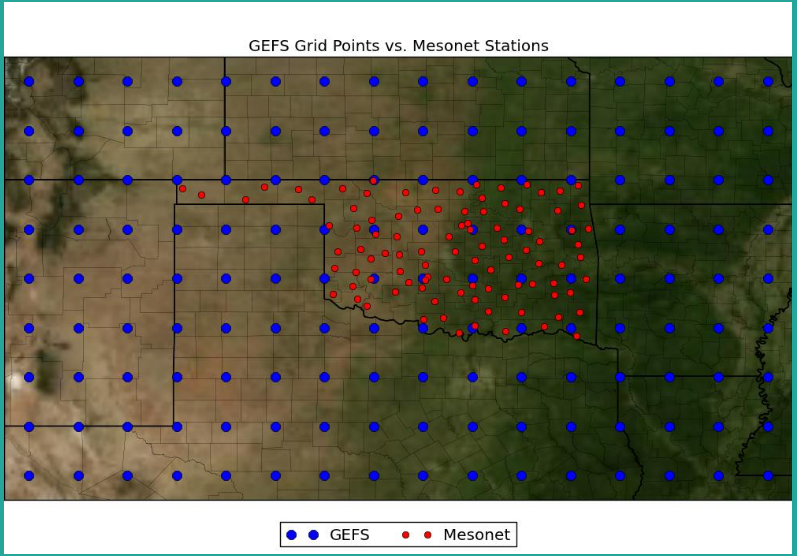
(CPU: Intel(R) Core(TM) i5-6200U CPU @ 2.30GHz)

Possible Extension of the Project

Higher Resolution Maps

Given maps with higher resolution could helps in the better understanding of the meteorological phenomena locally.

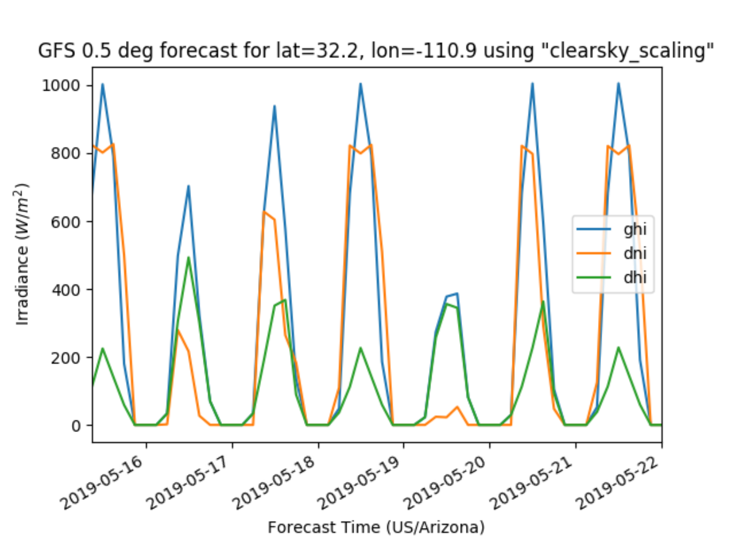
We can construct one more robust model including many solar parks that exist only few kilometers far one from the other



More Meteorological Data

Include data about:

- Pressure
- Total Cloud
- Humidity
- Wind Speed
- e.t.c



References

- ❖ ………… - Constantinos Theodorou
- ❖ Neural Networks for Data Science (Slides) - Simone Scardapane
- ❖ AMS 2013-2014 Solar Energy Prediction Contest - Kaggle
- ❖ Understanding Xavier initialization in DNN - Perpetual Enigma
- ❖ MultiDimensional Data – NetCDF, GRIB, HDF Format - Akshay Upadhyay