Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Bachelor Thesis

Title of the Thesis // Titel der Arbeit

Scientific software development of a Convolutional Neural Network for the reconstruction of missing data from a weather measurement station using numerical model data.

Wissenschaftliche Softwareentwicklung eines Convolutional Neuronal Networks für die Rekonstruktion fehlender Daten einer Wettermessstation unter Verwendung von Numerischen Modelldaten

---

Academic Degree // Akademischer Grad
Bachelor of Science (B.Sc.)

Author's Name, Place of Birth // Name der Autorin/des Autors, Geburtsort
Timo Wacke, Hamburg

Field of Study // Studiengang
Computing in Science (Physics Specialization)

Department // Fachbereich
Computer Science // Informatik

First Examiner // Erstprüferin/Erstprüfer
Prof. Dr. Thomas Ludwig

Second Examiner // Zweitprüferin/Zweitprüfer
Dr. Christopher Kadow

Matriculation Number // Matrikelnummer
7434883

Date of Submission // Abgabedatum
10.06.2024

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Eidesstattliche Versicherung

Wacke Timo
Last Name, First Name // Name, Vorname

Ich versichere hiermit an Eides statt, dass ich die vorliegende Abschlussarbeit mit dem Titel

Wissenschaftliche Softwareentwicklung eines Convolutional Neuronal Networks für die Rekonstruktion fehlender Daten einer Wettermessstation unter Verwendung von Numerischen Modelldaten

selbstständig und ohne unzulässige fremde Hilfe erbracht habe. Ich habe keine anderen als die angegebenen Quellen und Hilfsmittel benutzt sowie wörtliche und sinngemäße Zitate kenntlich gemacht. Die Arbeit hat in gleicher oder ähnlicher Form noch keiner Prüfungsbehörde vorgelegen.

Hamburg, den April 9, 2024
Place, Date, Signature // Ort, Datum, Unterschrift

Abstract

Zusammenfassung

Contents

List of Figures

1 Introduction

Weather station density varies greatly across the globe, depending on population density, economic development, and the availability of infrastructure. [3] While any weather station can experience downtime, the reliability of weather stations in regions with low station density is often low as well. So not only is downtime in regions where data is limited more likely but also more impactful because there are fewer neighboring stations to help compensate for the missing data. Not only would a denser more reliable network benefit weather forecasting, but it would also be beneficial for climate research. For example in East Africa, the weather station density is very low, but the region would be of great interest to the El Niño Southern Oscillation (ENSO) research. [1, 2] An innovative approach to increase the density of weather stations could be to use low-cost weather stations that could be 3D-printed and assembled by the local population. [2], either way, low-cost weather stations have reliability issues.

In light of the challenges posed by sparse weather station coverage, novel methodologies are required to address the reconstruction of missing weather data. One promising avenue involves the application of machine learning techniques, which offer a departure from traditional numerical reconstruction methods that are reliant on neighboring station data and are often computationally intensive. The application of machine learning in this case would be to connect numerical reanalysis data that is blurry and describes the weather in grid cells, with the local patterns that lead to measurements at a weather station. This would allow for independent operation and minimal computational resources needed for the appliance of the trained machine-learning model. By leveraging available local data, these techniques, such as Convolutional Neural Networks (CNNs), can be trained to estimate weather conditions at a designated time by assimilating global numerical weather model data. Despite the inherent blurriness of aerial data provided in grid cells, these models are anticipated to discern and adapt to local weather patterns such that they become capable of transferring knowledge from the meta situation to the local situation. The reanalysis of choice is the ERA5 reanalysis, which covers the globe in grid cells of 0.25° x 0.25°. The data is available in hourly timesteps from 1940 to the present and contains a wide range of variables, such as temperature, precipitation, wind speed, and many more.

To prove the concept it's likely easiest to start with temperature data, meaning the 2m temperature variable from the ERA5 reanalysis will be used as input to the neural net, one hour at a time, and the expected output will be the temperature at the weather station, during the same hour.

2 Conceptual Framework and Methodology

2.1 Data Acquisition and Preprocessing

The first step of the proposed methodology involves acquiring a dataset from a weather station. This dataset serves as the foundation for training and testing the Convolutional Neural Network (CNN). Upon obtaining the dataset, it is determined where temperature data is missing. While the weather station dataset is minute-based, data could be missing only for a few minutes within an hour instead of the full hour. To simplify the problem, the mean of the temperature values for each hour is calculated. And if all temperature values are missing for an hour, the hour is marked as missing. As explained in the introduction, the ERA5 reanalysis is hourly and covers the globe. The ERA5 data for the grid cells surrounding the weather station, at all the timesteps since the weather station data starts is obtained.
The ERA5 data then needs to be cropped to the neighboring grid cells, while centering the cutout as close to the weather station as possible. In the next step, the ERA5 data is divided into two datasets: one with all the hours marked as missing and one with all the hours marked as present. Until the model is trained, only the dataset with all the hours marked as present will be used.

2.2 Model Setup and Training

To determine if and to which extent the model learned to reconstruct the missing data, after we trained it the weather station dataset and the corresponding ERA5 dataset are split again into a pair of station and ERA5 data for training and one for validation. The training set is used to train the model, while the validation set is reserved to evaluate the model's performance. With the datasets prepared, the next phase involves configuring and training the Convolutional Neural Network (CNN) for the temperature reconstruction task. The CNN architecture is tailored to accept input in the form of 8x8 grid cells centered around the weather station's location. Employing a supervised learning approach, the CNN is trained using pairs of hourly temperature data from the weather station and corresponding grid cell data from ERA5. The training process iteratively feeds batches of data into the CNN, fine-tuning its parameters to minimize prediction errors and optimize accuracy in reconstructing missing temperature values.

2.3 Model Evaluation

Following the training phase, the CNN's performance is evaluated using the validation set. The model's capacity to accurately reconstruct missing temperature data at the weather station is scrutinized against ground truth values. This evaluation step serves to gauge the CNN's proficiency in capturing intricate weather patterns and producing precise temperature estimations. For that, the root mean squared error (RMSE) and the correlation coefficient are calculated. The RMSE is a measure of the differences between predicted and observed values, while the correlation coefficient quantifies the strength and direction of the linear relationship between the two datasets.

2.4 Application to New Data

Upon successful training and validation, the CNN is applied to the ERA5 dataset for the hours where station data is missing. These surrounding grid cell data from ERA5 are then fed into the trained CNN, which generates predictions for the missing temperature values at the weather station. Following the same approach as for validation the CNN reconstructs the weather data from the input grid cell data, as the model with its trained parameters is applied hour by hour. Thus the result is not a single continuous time series but a series of hourly predictions.

3 Theoretical Background

3.1 Convolutional Neural Networks

3.2 Reanalysis - ERA5

3.3 Weather Station Data Quality

4 Results

4.1 Results of basic setup

4.2 Experimenting with time context

5 Software Implementation

5.1 Copernicus Climate Data Store - CDS API

5.2 Data Preprocessing

5.3 CRAI - Climate Reconstruction AI

5.4 Infilling API

6 Discussion

6.1 How much data is needed?

6.2 Use for weather forecasting

References

[1] R. Marchant, C. Mumbi, S. Behera, and T. Yamagata. The indian ocean dipole—the unsung driver of climatic variability in east africa. *African Journal of Ecology*, 45:4–16, 2007. doi: 10.1111/j.1365-2028.2006.00707.x.

[2] R. Muita, P. Kucera, S. Aura, D. Muchemi, D. Gikungu, S. Mwangi, M. Steinson, P. Oloo, N. Maingi, E. Muigai, and M. Kamau. Towards increasing data availability for meteorological services: Inter-comparison of meteorological data from a synoptic

weather station and two automatic weather stations in kenya. *American Journal of Climate Change*, 10:300–303, 2021. doi: 10.4236/ajcc.2021.103014.

[3] Ariel Ortiz-Bobea. Climate, agriculture and food, 2021.