Adapting Transformer Encoder Architecture for Continuous Weather Datasets with Applications in Agriculture, Epidemiology and Climate Science

by

Adib Hasan

S.B. in Computer Science and Engineering and in Mathematics
Massachusetts Institute of Technology (2021)

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

MASTER OF ENGINEERING IN ELECTRICAL ENGINEERING AND COMPUTER SCIENCE

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

May 2024

Authored by:	Adib Hasan Department of Electrical Engineering and Computer Science May 17, 2024
Certified by:	Mardavij Roozbehani Principal Research Scientist of Laboratory for Information and Decision Systems, Thesis Supervisor
Certified by:	Munther Dahleh Professor of Electrical Engineering and Computer Science, Thesis Supervisor
Accepted by:	Katrina LaCurts Chair, Master of Engineering Thesis Committee

Adapting Transformer Encoder Architecture for Continuous Weather Datasets with Applications in Agriculture, Epidemiology and Climate Science

by

Adib Hasan

Submitted to the Department of Electrical Engineering and Computer Science
on May 17, 2024 in partial fulfillment of the requirements for the degree of

MASTER OF ENGINEERING IN ELECTRICAL ENGINEERING AND COMPUTER SCIENCE

ABSTRACT

This work introduces WeatherFormer, a transformer encoder-based model designed to robustly represent weather data from minimal observations. It addresses the challenge of modeling complex weather dynamics from small datasets, which is a bottleneck for many prediction tasks in agriculture, epidemiology, and climate science. Leveraging a novel pretraining dataset composed of 39 years of satellite measurements across the Americas, WeatherFormer achieves state-of-the-art performance in crop yield prediction and influenza forecasting. Technical innovations include a unique spatiotemporal encoding that captures geographical, annual, and seasonal variations, input scalers to adapt transformer architecture to continuous weather data, and a pretraining strategy to learn representations robust to missing weather features. This thesis for the first time demonstrates the effectiveness of pretraining large transformer encoder models for weather-dependent applications.

Thesis supervisor: Mardavij Roozbehani

Title: Principal Research Scientist of Laboratory for Information and Decision Systems

Thesis supervisor: Munther Dahleh

Title: Professor of Electrical Engineering and Computer Science

Acknowledgments

First and foremost, I am deeply grateful to my parents for their unwavering support throughout my MEng journey. Their encouragement and belief in me have been a constant source of strength. I would also like to express my heartfelt thanks to my grandmother and aunts for their endless love, emotional support, and words of encouragement.

I am sincerely thankful to my friends, who have not only offered guidance and support but also challenged me intellectually, pushing me to grow beyond my perceived limits.

I owe a special debt of gratitude to Mardavij for his patient guidance and insightful answers to my countless questions. His expertise has been invaluable to my research. I would also like to express my gratitude towards Mark and David, who have offered insights and guidance during different times of the research. Additionally, I am thankful to Munzer for providing me with the opportunity to work in a wonderful lab environment surrounded by so many brilliant minds.

Finally, I would like to acknowledge the MIT community and the clubs I participated in for making me feel at home during my time away from home.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Understanding the changing weather patterns is extremely important in several major fields including agriculture, epidemiology, climate science, disaster response, and transportation. However, real-world datasets in these fields are small and lack detailed weather measurements. For instance, a crop yield prediction dataset often contains only 5-7 years of detailed data for a few farms [McFarland et al., 2020] or just a few weather measurements over a long period [Khaki et al., 2020]. Consequently, large models with sufficient capacity to learn weather patterns overfit when trained on these datasets. On the other hand, without a good representation of weather, any model's performance for prediction tasks in these fields will remain sub-optimal.

In Natural Language Processing (NLP), the problem of small datasets is tackled by training a large model on a massive unlabelled text dataset, such as the English Wikipedia [Devlin et al., 2019], and then finetuned on small datasets for prediction tasks. This approach with pretrained large models like BERT [Devlin et al., 2019], RoBERTa [Liu et al., 2019b], and others [Clark et al., 2020, Lan et al., 2020] have demonstrated remarkable success in improving the benchmarks on sentiment analysis [Batra et al., 2021], machine translation [Zhu et al., 2020], and reading comprehension [Fernandez et al., 2023].

Unlike text, which is a discrete domain, weather data is continuous and has both spatial and temporal dependencies. Hence, it is unclear if large models pretrained on public weather datasets will improve the performance of the downstream tasks. While researchers, such as Man et al. [2023] and Nguyen et al. [2023], have proposed transformer decoder-based [Vaswani et al., 2017] weather models for weather forecasting and downsampling, no foundational

weather model to our knowledge has been trained to extract good representations of weather from a small number of observations.

In this thesis, we aim to fill this knowledge gap by pretraining a weather *encoder model*, called WeatherFormer, on a large dataset of satellite-based weather measurements from the NASA Power Project [NASA, 2024]. We also show that finetuning this model for yield prediction and influenza forecasting achieves state-of-the-art performance in both tasks. We lastly show that Masked Language Modeling (MLM), a common pretraining strategy for textual data, is ineffective for pretraining models in the continuous weather domain, and propose a novel pretraining strategy to address this shortcoming.

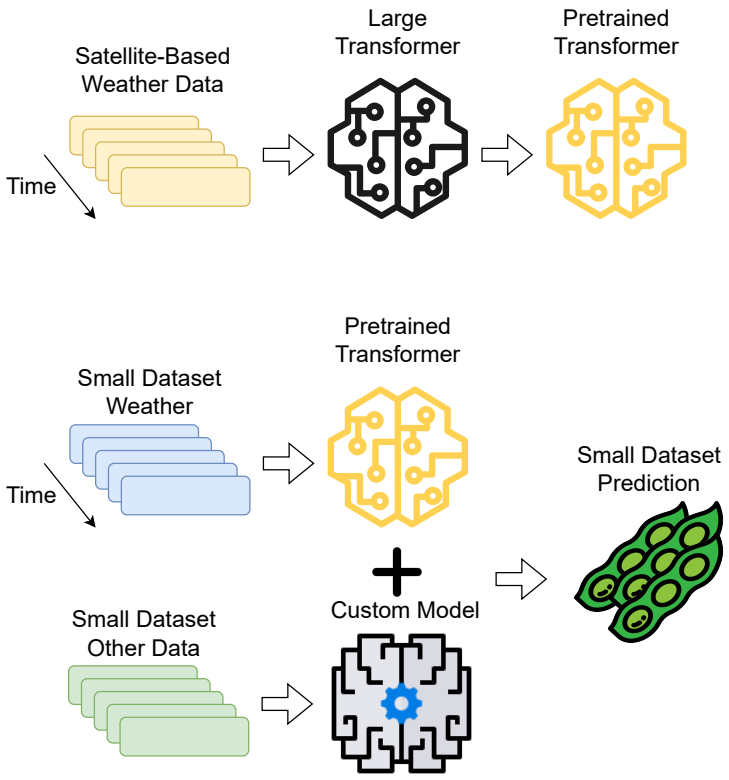


Figure 1.1: A large transformer model (WeatherFormer) is pretrained on a satellite-based massive weather dataset, enabling the model to learn rich representations of weather during pretraining. This pretrained model can extract robust weather features for a new prediction task (Small Dataset) even when only a limited number of weather measurements are available for that task. The learned weather features can be used to improve the prediction accuracy for the new task.

1.2 Proposed Contribution

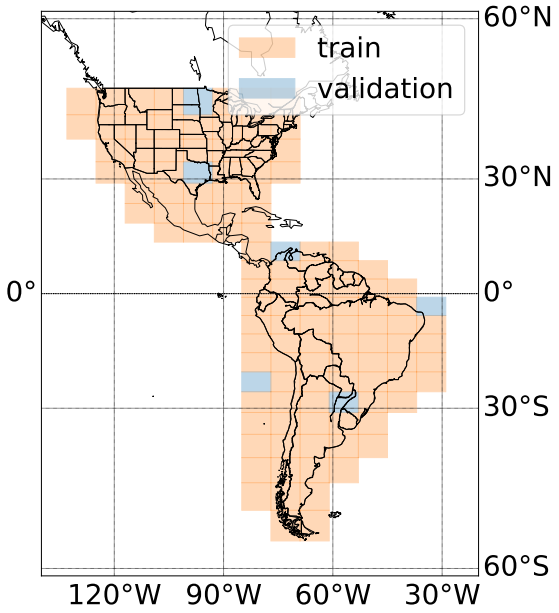


Figure 1.2: Map of the pretraining data. 5% of the grid rectangles were selected at random for validation and the rest were used for pretraining the model.

Our proposed foundational weather model, WeatherFormer, is a transformer encoder-based model. It is not trained to predict the weather in the future, but rather to extract good representations of weather in a small dataset, and with just a few basic measurements such as the mean temperature and precipitation.

We trained WeatherFormer on a novel pretraining task and a large pretraining dataset, allowing the model to learn good representations of weather. It also incorporates a new positional encoding mechanism sensitive to geographical location, year, and seasonality, enabling it to capture the dynamic and repetitive nature of weather patterns across different times and places. It supports a maximum sequence length of 365, allowing it to process 1 year of daily weather data, 7 years of weekly weather data, and 12 years of monthly weather data.

The pretraining dataset of WeatherFormer consisted of 31 satellite-based daily weather measurements over the continental United States, Central America, and South America for 39 years. (Figure 1.2) 5% of the grids were chosen at random for validation and the remaining data was chosen for pretraining the model. This large pretraining dataset allows the model to extract weather representations even if only a few (1-6) weather measurements are present in the downstream datasets.

The raw satellite data was downloaded from NASA Power Project [NASA, 2024] and was then enhanced with estimates of additional variables derived from meteorological equations. The model was trained to predict 10 weather measurements from 21 input weather measurements. In every batch during training, some input and output variables are swapped so that the model learns the relationship among all variables.

Our experiments show that the pretraining process enhances the model's performance in predicting county-level crop yields and influenza forecasting. Similar to the impact seen in NLP, pre-training on a large dataset allows the model to grasp the complexities of weather, leading to marked improvements in task-specific applications.

Our contributions can be summarized as follows:

- Collected, processed, and open-sourced a 60 GB dataset of satellite weather measurements ready for training large deep learning models.
- Modified the transformer encoder model with a novel positional encoding, scalers, and a novel pretraining task to pretrain on a large volume of continuous weather data.
- Trained and open-sourced two foundational models for weather, called WeatherFormer, with 2 million and 8 million parameters, respectively.
- Finetuned WeatherFormer models to achieve SOTA performance in Soybeans yield prediction in the US corn belt and Influenza forecasting in New York City.

1.3 Applications

Given WeatherFormer's ability to find good representations of weather from a limited amount of data, we expect the model to be useful in a variety of domains. Some potential future applications include:

- **Agriculture**: Most agricultural datasets contain either a few weather measurements [Khaki et al., 2020], or less than ten years' worth of detailed measurements with many missing entries [McFarland et al., 2020]. As a result, large deep learning models cannot be trained on such datasets. We believe finetuning WeatherFormer-based models can be one approach to improving the SOTA performance for agricultural tasks. We empirically demonstrate this for county-level soybean yield prediction with the Khaki et al. [2020] dataset. Additional agricultural applications could be predicting plant

disease epidemics, biomass accumulation, and flowering periods, all of which are very important problems to solve for data-driven and resource-efficient agriculture.

- **Epidemiology:** Several diseases are known to be influenced by weather parameters. For instance, average mean temperature, humidity, and wind speed affect influenza outbreaks [Amendolara et al., 2023]. Dengue transmission is influenced by rainfall, relative humidity, and minimum temperature. [Abdullah et al., 2022] Similar to dengue, malaria epidemics are influenced by weather conditions that promote the breeding of Anopheles mosquitoes. [Hoshen and Morse, 2004] Cholera and typhoid outbreaks are affected by weather variables and climate change [Christaki et al., 2020, Jia et al., 2024] WeatherFormer can be finetuned to improve the prediction of the outbreaks of these diseases, allowing better disease control and prevention, which can save thousands of human lives. We validate this claim by forecasting weekly influenza cases in New York City with WeatherFormer and achieving the best performance among all tested models.

- **Climate Science:** Weather patterns are critical drivers of environmental phenomena such as droughts, coastal floods, wildfires, soil erosion, pollution, and air quality. Due to the changing climate, weather patterns are undergoing a major shift, modifying the frequency and intensity of such events. For instance, drought conditions in the northern hemisphere are expected to worsen, with increasing frequency and intensity [Balting et al., 2021]. Additionally, projections indicate that by 2100, 52% of the global population could be at risk of coastal flooding due to rising sea levels and extreme weather events [Kirezci et al., 2020]. Soil erosion is also predicted to escalate, particularly in semi-arid regions [Eekhout and de Vente, 2022]. A foundational model like WeatherFormer can be finetuned to predict the expected changes to the environment, aiding in climate change research.

Chapter 2

Background

2.1 Transformer Architecture

Transformer models have revolutionized the field of natural language processing (NLP) since their introduction by Vaswani et al. [2017]. The core idea behind transformers is to model relationships between all parts of the input sequence, regardless of their positional distances. This is achieved through the attention mechanism, which computes the attention scores based on the query (Q), key (K), and value (V) matrices derived from the input embeddings. The attention function can be described mathematically as follows:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{2.1}$$

where d_k is the dimension of the key. By stacking layers containing a self-attention mechanism and a position-wise fully connected feed-forward network, transformers can capture complex dependencies with remarkable efficiency. Since the self-attention mechanism is position-agnostic, a positional encoding is added to the input before the self-attention. Vaswani et al. [2017] proposed the sinusoidal positional encoding for this as shown below:

$$PE_{\text{pos},2i} = \sin\left(\frac{\text{pos}}{10000^{2i/d}}\right)$$
$$PE_{\text{pos},2i+1} = \cos\left(\frac{\text{pos}}{10000^{2i/d}}\right)$$

where d is the hidden dimension and $0 \leq i < d$.

2.2 Foundational Models in Deep Learning

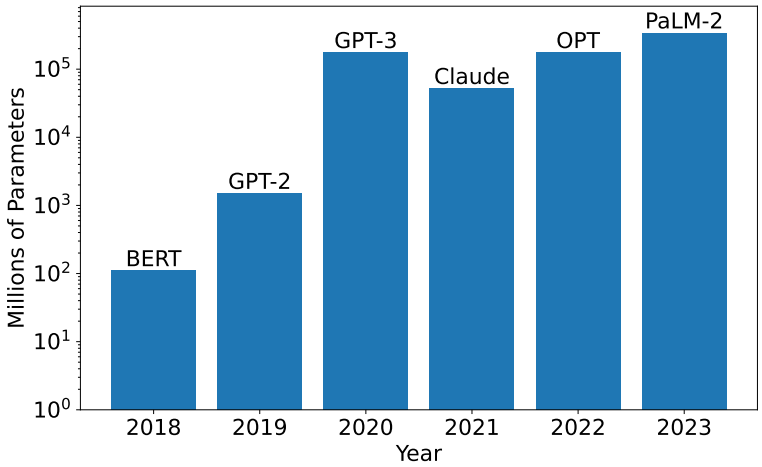


Figure 2.1: In Natural Language Processing, foundational models (also called *Large Language Models (LLMs)*) have grown in the number of parameters in the recent years.

Foundational models, exemplified by GPT [Radford et al., 2018] and BERT [Devlin et al., 2019], mark a transformative shift in the development and deployment of deep learning systems in Natural Language Processing (NLP). These models are pre-trained on extensive data corpora and demonstrate an exceptional capacity to generalize across diverse tasks without task-specific training.

The foundational models have also found applications beyond NLP. In biology, Le et al. [2022] demonstrated that pre-training a BERT-like model on DNA sequences enhances the prediction accuracy for DNA promoters. In the medical field, models that integrate text and imagery have shown promise [Azad et al., 2023].

In meteorology, foundational models have been introduced for short and medium-term weather prediction, downsampling, and climate-related text analysis. Pathak et al. [2022] developed a time series-based recurrent neural network [Hochreiter and Schmidhuber, 1997] for short to medium-range weather forecasting. Lam et al. [2023] employed Graph Neural Networks [Scarselli et al., 2009] to process satellite imaging data for global medium-range weather predictions. More recently, transformer-based models have been deployed for weather forecasting [Man et al., 2023, Nguyen et al., 2023] and data downsampling [Nguyen et al., 2023]. Lastly, researchers have developed both encoder-based [Fard et al., 2022, Webersinke et al., 2022] and decoder-based models [Bi et al., 2024] for analyzing climate-related textual data. To the best of our knowledge, no foundational weather model besides WeatherFormer

has been trained to extract weather representations for downstream prediction tasks.

2.3 Self-Supervised Learning

The pretraining datasets used to train large foundational models often come with no labels. Hence, the models cannot be trained on the datasets with traditional loss functions such as Mean Square Error (MSE) or Negative Log Likelihood (NLL). Instead, the researchers have introduced new learning tasks that allow the models to learn information from the dataset without labels. This approach to learning is called Self-Supervised Learning. Below we describe some of the common learning tasks used in natural language processing. For each of them, suppose $X = (x_1, x_2, \ldots, x_n)$ is a sentence with n tokens and Θ is the collection of the model parameters.

- **Masked Language Modeling (MLM):** Originally introduced by BERT, MLM involves masking a fraction of input tokens, which are then predicted by the model based on the remaining tokens. Suppose $\mathcal{M}$ is the set of masked tokens and $\mathcal{V}$ is the set of visible tokens to the model. The loss function for MLM can be written as:

$$\mathcal{L} = -\log P(\mathcal{M}|\mathcal{V}, \Theta)$$

- **Autoregressive Language Modeling**: Used by models like GPT [Brown et al., 2020], this task involves predicting the next word in a sequence given the previous words. This trains the model to understand the probability distribution of a language and is useful for generating coherent text sequences. Mathematically, this task can be written as:

$$\mathcal{L} = -\sum_{k>1}^{n} \log P(x_k|x_1, \ldots, x_{k-1}, \Theta)$$

  A variation of this task is also effective for continuous weather data, as confirmed by Nguyen et al. [2023].

- **Replaced Token Detection:** As used in ELECTRA [Clark et al., 2020], this task involves one generator and one discriminator network. The generator network replaces some tokens from the input with some 'fake' tokens, and the discriminator predicts whether each token was replaced by the generator or not. If $\hat{X}$ is the input sentence

after modification, then the discriminator's loss function is:

$$\mathcal{L} = -\sum_{k=1}^{n} \log P(x_k = \text{real}|\hat{X}, \Theta)$$

This task was reported to be more efficient for pretraining than MLM, especially for the small models. [Clark et al., 2020]

- **Next Sentence Prediction:** Also introduced by BERT, this technique trains models to determine the correct order of two segmented inputs, typically sentences. Liu et al. [2019b] reported that this task was unnecessary and MLM alone provided sufficient pretraining with a larger training dataset.

Chapter 3

Architecture and Design Choices

In this chapter, we describe the architecture of the WeatherFormer in detail. Figure 3.1 shows a visualization of the forward pass through the architecture.

3.1 Limitations of the Transformer Encoder Models

Transformer encoder models like BERT have been widely successful in natural language processing due to their ability to capture intricate dependencies within discrete textual data. However, adapting such architectures to continuous domains like weather data presents additional challenges. While the success of Vision Transformer (ViT) models [Dosovitskiy et al., 2021] demonstrates the applicability of transformers to images, weather patterns exhibit spatiotemporal variations that necessitate designing custom positional encodings.

Another challenge arises from the potential discrepancy between the pretraining and fine-tuning datasets, where the available weather measurements and their temporal granularities may vary. Hence, the model needs to be able to accept partial observations as inputs. This makes pretraining challenging with MLM since no feature is missing across the entire input sequence during pretraining, making the model go out of distribution during the fine-tuning phase.

Furthermore, the self-attention mechanism inherent to transformers has a quadratic computational complexity ($O(N^2)$) with respect to the input sequence length (N). Hence, the context length needs to be limited.

3.2 Our Design

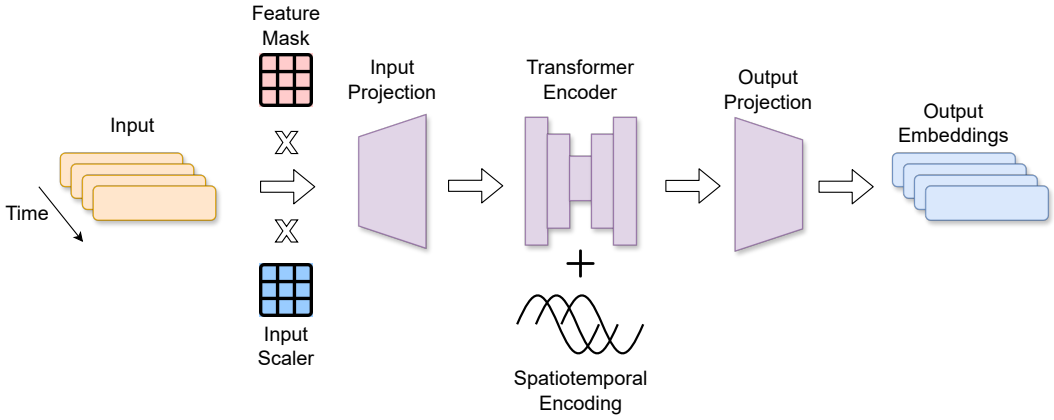


Figure 3.1: The forward pass of the weather inputs through the WeatherFormer architecture. The input is first multiplied with learnable input scalers and a feature mask and then projected to a hidden dimension through a linear layer. After that, the input goes through a transformer encoder with a novel spatiotemporal encoding mechanism and finally, the input is projected to an output dimension.

3.2.1 Feature Mask and Padding Mask

The WeatherFormer model expects an input of the shape $N \times 31$, where N is the length of the input sequence and 31 is the number of weather measurements. However, during pertaining and task-specific fine-tuning, not all of the input weather variables may be present. For this reason, WeatherFormer also accepts a weather mask. When this mask is true for a weather measurement, the corresponding weather measurement is considered to be missing and filled in with 0. The WeatherFormer can accept variable length input sequences with a padding mask, however, the maximum context length is 365. Therefore, the model can process a maximum of 1 year of daily data, 7 years of weekly data, and 12 years of monthly weather measurements. For each input in the sequence, the WeatherFormer will output a vector representation of the input.

3.2.2 Scaling Parameters

To adapt transformer architectures, traditionally used for discrete NLP tokens, to continuous weather data, we introduce learnable scaling parameters for each weather input dimension. This is essential as standardized weather measurements (mean 0, std. dev. 1) might not align

with the fixed range of the sinusoidal positional encodings ($[-1, 1]$), potentially hindering optimal representation. Our scaling parameters address several challenges:

- Ensures that positional encodings contribute meaningfully to the learned representations of weather data.
- Enables the model to differentially rank the weather measurements, reflecting their varying importance in determining the overall weather conditions.
- Integrates temporal scales (daily, weekly, monthly) directly into the model architecture through distinct scaling embeddings.

We implement scaling with a PyTorch Embedding layer, assigning different embeddings for temporal granularity of input between 1 and 30. All embeddings are initialized with 1 and the embeddings for 1, 7, and 30 (corresponding to daily, weekly, and monthly input data) are learned during pretraining.

3.2.3 Spatiotemporal Positional Encoding

The attention mechanism in the transformer does not take the position of a token into account. For this reason, an encoding for position is added to the input. However, in the case of WeatherFormer, the input is a sequence of weather measurements over some time, which has both temporal and spatial dependency. The original sinusoidal positional encoding of the transformers will be unable to take this into account. For this reason, we designed a new Spatiotemporal Positional Encoding for the WeatherFomer.

Our key insights come from the following two observations:

- Weather shows an annual repeating pattern, however over longer periods there are repeating patterns like El Niño [Philander, 1989] and drifts due to climate change.
- Weather also demonstrates spatial variability. So a positional encoding for weather input needs to encode the latitude and the longitude as well.

By combining these two insights, we came up with the following positional encoding for

weather:

$$
\begin{aligned}
PE(\text{year}, \text{lat}, \text{lng})_{\text{pos},4i} &= \sin\left(\text{pos} \cdot \cdot 10000^{-4i/d}\right) \\
PE(\text{year}, \text{lat}, \text{lng})_{\text{pos},4i+1} &= \cos\left(\text{pos} \cdot \cdot 10000^{-4i/d}\right) \\
PE(\text{year}, \text{lat}, \text{lng})_{\text{pos},4i+2} &= \sin\left(\frac{\pi \cdot \text{lat}}{180} \cdot \cdot 10000^{-4i/d}\right) \\
PE(\text{year}, \text{lat}, \text{lng})_{\text{pos},4i+3} &= \cos\left(\frac{\pi \cdot \text{lng}}{180} \cdot \cdot 10000^{-4i/d}\right)
\end{aligned}
$$

The scaler $\pi/180$ converts the coordinates into radians and also ensures in every 360° distance, the spatial encoding repeats itself.

3.2.4 Transformer Encoder and Output Projection

Once both the temporal granularity encodings and the spatiotemporal encodings are added to the input, the input is then passed through a transformer encoder. After the transformer encoder, the model is finally projected to an output dimension. For an input of the shape $N \times 31$, the model produces an output of the shape $N \times M$, where M is the desired output dimension.

Chapter 4

Pretraining

4.1 Weather Data Collection

Weather exhibits complex spatiotemporal patterns that are not readily evident from a small dataset. Additionally, not all of the important weather measurements are available for specific downstream tasks. Therefore, a foundational model of the weather should be able to create a good representation of the weather from only partial observations in the downstream tasks. For this reason, we chose to pretrain the weather model on a large-scale satellite dataset. Satellite data is globally available, whereas the data from local weather stations may not measure the same weather variables across the globe.

Our pretraining data was downloaded from the NASA Power API [NASA, 2024]. Although the Power API has kept records of weather since 1980, we have found that the earlier years were missing many of the important weather measurements and we chose to skip them. We downloaded 28 daily weather measurements from 1984 till 2022. The dataset consisted of rectangular regions of shape $5° \times 8°$ spanning the continental United States, Central America, and South America as shown in Figure 1.2. The dataset contained 119 rectangles, each containing 160 unique coordinates with weather measurements. The spatial resolution of the dataset was 0.5 degree^2.

The dataset contained a small percentage of missing values, which were subsequently imputed using data from preceding years. A complete list of all the weather measurements is given in Table A.1.

Additionally, we also computed weekly and monthly averages of the weather variables and added them to the datasets. Consequently, the WeatherFormer is trained on daily,

weekly, and monthly granularities. The pretraining dataset in total had approximately 9.1 billion floating point numbers. We used 95% of the data for pretraining and 5% data for validation.

4.2 Additional Measurement Estimation

Some weather measurements, such as evapotranspiration, vapor pressure difference, etc. are important for agriculture, epidemiology, etc but were not available in the downloaded data. Hence they were estimated using meteorological equations as described below.

4.2.1 Tetens Equation

The Tetens equation, as described by Tetens [1930], offers a straightforward method to estimate the saturation vapor pressure (e_a) over liquid water or ice, contingent on the ambient temperature. For temperatures above freezing ($x > 0°C$, representing liquid water), the equation is:

$$e_a = 0.6108 \cdot \exp\left(\frac{17.27 \cdot x}{x + 237.3}\right)$$

Conversely, for temperatures at or below freezing ($x \leq 0°C$, representing ice), the equation modifies to:

$$e_a = 0.6108 \cdot \exp\left(\frac{21.87 \cdot x}{x + 265.5}\right)$$

In these equations, e_a represents the saturation vapor pressure in kPa, and x denotes the temperature in degrees Celsius ($°C$).

4.2.2 FAO Penman-Monteith Equation

Accurate estimation of evapotranspiration (ET) is crucial for water resource management, agricultural planning, and understanding the hydrological cycle. The FAO Penman-Monteith equation [Ndulue and Ranjan, 2021] is widely accepted as the standard method for calculating reference evapotranspiration ($ET0$) from climatic data. The equation is expressed as:

$$ET0 = \frac{0.408\Delta(R_n - G) + \gamma\frac{900}{T+273}u_2(e_s - e_a)}{\Delta + \gamma(1 + 0.34u_2)}$$

where:

- $ET0$ is the reference evapotranspiration (mm day^{-1}),
- Δ is the slope of vapor pressure curve (kPa $^{\circ}\text{C}^{-1}$),
- R_n is the net radiation at the crop surface (MJ m^{-2} day^{-1}),
- G is the soil heat flux density (MJ m^{-2} day^{-1}),
- T is the mean daily air temperature at 2 m height ($^{\circ}$C),
- u_2 is the wind speed at 2 m height (m s^{-1}),
- e_s is the saturation vapor pressure (kPa),
- e_a is the actual vapor pressure (kPa),
- γ is the psychrometric constant (kPa $^{\circ}\text{C}^{-1}$).

This equation assumes a standard grass reference crop with an assumed height of 0.12 m, a fixed surface resistance of 70 sm^{-1}, and an albedo of 0.23.

4.3 Masked Feature Prediction

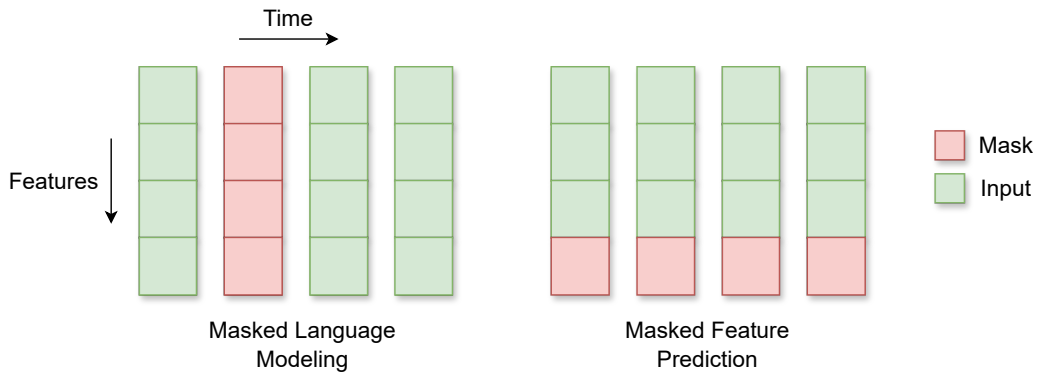


Figure 4.1: A comparison between Masked Language Modeling in NLP vs Masked Feature Prediction in WeatherFormer. For the latter, some input features across the time dimension are dynamically masked during pretraining. This task naturally allows the model to handle missing features during finetuning.

Due to the continuous nature of the weather data and the discrepancy between pretraining and finetuning features, a majority of the standard pretraining tasks used in NLP cannot be directly used for weather data. For instance, the MLM task cannot handle missing features during the fine-tuning phase, which makes the model go out of distribution. Similarly, the autoregressive language modeling is not directly applicable for WeatherFormer since the objective is not to predict weather in the future, but instead to learn from limited finetuning datasets. Replace token detection also does not work very well for this task since weather data is continuous and every value has a meaning.

After considering the pros and cons of various pretraining tasks, we decided to design a new pretraining task based on MLM. We call this task *Masked Feature Prediction (MFP)*.

In this task, 10 weather measurements across the sequence length were replaced with zeros, and the model predicted them from the remaining 21 weather features with MSE loss. In each batch during training, one input and one masked variable are swapped. Thus throughout the training, the model learns the relationship between every pair of weather measurements.

This pretraining approach addresses the inherent challenge of having few weather measurements available in real-world prediction scenarios. By training the model to predict missing weather measurements on a pertaining dataset, we encourage the development of robust representations derived from partial observations, making it well-suited for our downstream tasks.

4.4 Optimization

We pretrained two models of size 2M and 8M, respectively.

Both models were pretrained for 75 epochs over the training data and computed performance over the validation dataset after every epoch. We tested with 16, 21, and 26 input features for MFP, and during our downstream task validation, the model with 21 input features performed the best.

Our loss function was MSE and Adam [Kingma and Ba, 2017] was used as the optimizer with a learning rate of 0.0005, a warm-up period of 10 epochs, a batch size of 64, and an exponential learning rate decay factor of 0.99. These hyperparameters were selected after pretraining small-scale models during other preliminary investigations. The 2M parameter model took 17 hours to train on 2 NVIDIA Volta 100 GPUs and the 8M parameter model took 39 hours. We did not explore the full search space for the full-scale models due to the prohibitive expansiveness of the task.

Chapter 5

Applications: Crop Yield Prediction

5.1 Background

Crop yield prediction is an important task for ensuring global food security, managing agricultural resources, and supporting the development of new crop hybrids. However, the task is difficult due to the variability in weather patterns, soil conditions, and the scarcity of comprehensive public datasets.

Additionally, different crops have different growing periods and development stages. This diversity necessitates a tailored approach for each crop type, making it very difficult to create a universal yield prediction model. Previous research in crop yield prediction has focused on two different types of models as described below.

5.2 Simulation Based Models

As the name suggests, the simulation-based models are based on principles of physics and plant science to simulate crop growth and forecast yields. The models are crop-specific, reflecting the distinct growth patterns and environmental responses of different crops. Some notable examples include SIMPLE [Zhao et al., 2019], WOFOST [van Diepen et al., 1989], and APSIM [Keating et al., 2003], all of which mimick various growth phases of plants with varying degrees of complexities. For reference, we describe the SIMPLE model below.

SIMPLE Model

The SIMPLE model is a mechanistic model that models the biomass growth of a plant under varying environmental stress factors. Here's a simplified set of equations that represent this model

$$\text{Biomass}_{\text{rate}} = \text{Radiation} \times f\text{Solar} \times \text{RUE} \times f(\text{CO}_2) \times f(\text{Temp}) \times f(\text{Env.Stress}) \quad (5.1)$$

$$\text{Biomass}_{\text{maturity}} = \sum_{\text{days}} \text{Biomass}_{\text{rate}} \quad (5.2)$$

$$\text{Yield} = HI \times \text{Biomass}_{\text{maturity}} \quad (5.3)$$

where fSolar is the fraction of the radiation intercepted by the plant canopy, RUE is the radiation use efficiency of the plant, HI is the harvest index, and $f(\text{CO}_2)$, $f(\text{Temp})$, $f(\text{Env.Stress})$ are various non-linear functions designed to capture environmental effects on plant growth.

It is important to note that despite their scientific basis, simulation-based models require extensive calibration for accuracy. Even with meticulous fine-tuning, the results often fall short of being highly reliable, particularly due to their dependence on detailed, test site-specific data. [Kahraman, 2021]

5.3 Machine Learning-based Yield Prediction

Many researchers have explored various machine learning-based approaches to predict crop yields, including those of wheat, corn, soybeans, rice, and potatoes. Data sources for these studies range from on-field environmental and soil data Ruß and Kruse [2010] to satellite-based measurements Khaki et al. [2020]. Farm-level data typically provides rich and accurate features, including weather measurements, soil characteristics, fertilizer applications, irrigation details, and pest control information [Ahamed et al., 2015]. In contrast, satellite-based data often lacks this granularity and accuracy, making it less suited for detailed predictions. Consequently, satellite-based data are generally used for broader yield predictions at the county, state, or national level, where the effects of individual agricultural practices are less pronounced. [Khaki et al., 2020]

Commonly employed algorithms in this field include artificial neural networks, regression models, random forests, clustering algorithms, and ensemble models. Artificial neural networks [Ruß et al., 2008] and regression models [Ruß and Kruse, 2010] have been successfully applied to predict farm-level wheat yields in Germany. Clustering algorithms have been uti-

lized by Ahamed et al. [2015] to forecast the yields of rice, potatoes, and wheat in Bangladesh. A hybrid model combining convolutional [LeCun et al., 1998] and long short-term memory [Hochreiter and Schmidhuber, 1997] neural networks has been used to predict county-level corn and soybean yields in the United States based on satellite-derived weather and soil data Khaki et al. [2020]. Lastly, Jeong et al. [2016] has developed a random forest framework for predicting regional and global yields of corn, potatoes, and wheat.

5.4 Yield Dataset

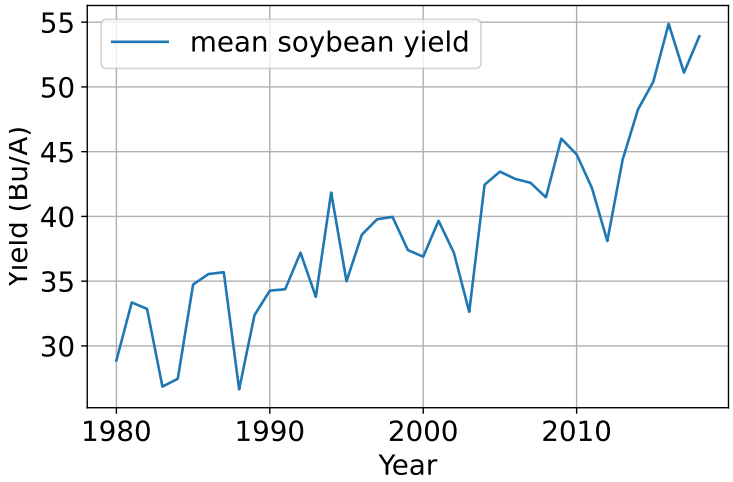


Figure 5.1: Mean soybean yield (Bu/A) across 9 corn belt counties of the United States. The yield gradually increased due to hybrid vigor and better farming practices.

To evaluate WeatherFormer's performance on yield prediction, we used a soybean yield dataset introduced by Khaki et al. [2020]. This dataset contains county-level soybean yield for the nine states in the corn belt region of the United States from 1980 to 2018. Alongside soybean yield data, it also contains six weather variables (precipitation, solar radiation, snow water equivalent, maximum and minimum temperatures, and vapor pressure), soil profile across six depths for 10 characteristics (composition, bulk density, and organic matter content, etc) and management practices. Notably, the original paper reported results for 13 states, but the publicly available dataset comprises data for only 9 states. Weather and practice data are provided weekly, whereas soil data lack a temporal component. The experiment was run five times and during each run, the data from randomly chosen seven states were used for training and the remaining two states were used for validation. The best Root Mean Square Error (RMSE) for the validation dataset from each of the five runs is averaged and reported.

It is important to note that the weather measurement source and granularity in this dataset are different from the pretraining data for WeatherFormer, and despite that the model showed considerable improvement over the baseline, suggesting a good understanding of the weather from pretraining.

5.5 Finetuning WeatherFormer for Yield Prediction

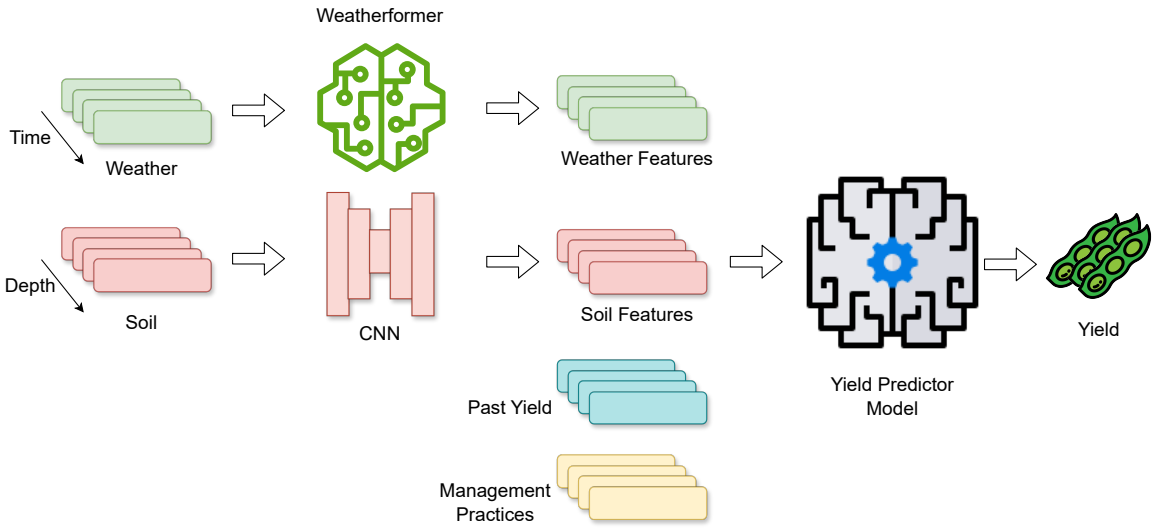


Figure 5.2: Soybean yield predictor architectures utilizing WeatherFormer. The weather measurements are processed with WeatherFormer and the soil measurements are processed with a CNN by Khaki et al. [2020]. Then the yield is predicted with either a linear layer or a transformer. The entire model is trained at once. Since yield for the current year is the target variable, it is replaced in the input with last year's yield.

WeatherFormer was pretrained on a large weather dataset to learn a good representation of the weather from a few measurements. However, crop yield does not depend on weather alone. Soil profile, cultivation practices, hybrid vigor, and fertilizer use also affect the yield. Therefore, we decided to build hybrid models for yield prediction, where the weather is processed by WeatherFormer and the remaining data is processed by models published by other researchers such as Khaki et al. [2020], and then finally the yield is predicted. The entire model is trained together.

We tested two models with WeatherFormer components. Both of them are depicted in Figure 5.2. Below we describe the models.

- **WeatherFormer+Linear Model:** Weather features are processed with Weath-

erFormer and soil properties are processed with the CNN used in Khaki et al. [2020]. These features are concatenated with management practices and past yields. A single linear layer predicts yield from this input, highlighting WeatherFormer's direct impact. This setup isolates WeatherFormer's contribution, but it is suboptimal for yield forecasting since soybean yields show clear trends over time as shown in Figure 5.1. These trends are not captured by a simple linear layer in this setup.

- **WeatherFormer + Transformer:** Like before, weather features are processed by the WeatherFormer, and soil features are processed by the CNN. These processed features, along with management practices and past yield, are passed through a transformer to capture temporal dependencies and then yield is predicted. The second transformer had a hidden dimension of 64, 8 heads, and 3 layers. The WeatherFormer + transformer architecture aligns with the time-dependent nature of the task and is optimized for the best performance in yield prediction.

5.6 Baseline Models

We evaluated three baseline models: a least squares linear regression model, the CNN-RNN model as proposed by Khaki et al. [2020], and a novel CNN-Transformer model.

The CNN-RNN model leverages weather data, soil characteristics, management practices, and past mean yield across all training counties (Figure 5.1) to predict crop yields effectively. Initially, this model employs two separate CNNs to extract features from soil data across the depth dimension and weather data across the time dimension for the last three years. These features are then concatenated with the remaining data and passed through an LSTM to predict the yield for the current year. We also observed that instead of the mean yield, the past yield for a location is a better predictor of future yield and the CNN-RNN model does considerably better with past yield as input. Since the yield for the current year is the target variable, it was replaced by the yield from the most recent year in the inputs.

While the CNN-RNN model effectively utilizes environmental, management, and temporal data for yield prediction, it does not account for spatial variations in yield. To address this limitation, we modified the model by replacing its LSTM component with a transformer encoder, allowing the mode to use our novel spatiotemporal encoding described before. This transformer also had a hidden dimension of 64, 8 heads, and 3 layers. This model, referred to as the *CNN-Transformer* model, has demonstrated state-of-the-art (SOTA) performance barring the WeatherFormer models on this soybean yield prediction dataset and is a

good competitor architecture to the WeatherFormer-based models.

5.7 Optimization and Training

In our experiments, we optimized our models using the Adam optimizer with an initial learning rate of 0.0005, applying a 10-epoch warm-up period followed by an exponential decay with a factor of 0.95. We conducted training over 40 epochs, using a batch size of 64, and assessed model performance on the validation set using the root mean square error (RMSE).

We thoroughly explored hyperparameter sensitivity, testing learning rates ranging from 0.0004 to 0.0012, warm-up periods from 4 to 10 epochs, and batch sizes of 32, 64, and 96. We discovered all models are the most sensitive to the temporal depth of the input data. Specifically, the CNN-RNN architecture demonstrated optimal performance when trained with 3 years of historical data, whereas the transformer-based models achieved the best results with 7 years of data, suggesting differing capacities for capturing long-term dependencies.

To accommodate various data histories, we trained all models using 1 to 7 years of past weather data. The training procedures were executed on an NVIDIA Volta 100 GPU, typically requiring between 1 to 2 hours depending on the model complexity and data span.

Chapter 6

Applications: Influenza Forecasting

6.1 Background

Influenza viruses are responsible for recurring annual epidemics worldwide, particularly impacting temperate regions. The severity and infectiousness of influenza vary each season due to the virus's rapid mutation rate and weather. The Centers for Disease Control and Prevention (CDC) reported that there have been more than 380,000 hospitalizations and 24,000 deaths from influenza during the 2023-2024 season alone. [CDC, 2024]

To address these challenges, researchers have developed a variety of forecasting models. These models can be broadly categorized into two types as described below.

6.2 Parametric Models

The first category includes parametric models that utilize disease dynamics and epidemiological data to model influenza. These models apply statistical learning techniques to fit time-series data of influenza cases. The first seminal work in this field was the Susceptible-Infected-Retired (SIR) model [Kermack and McKendrick, 1927], which is the backbone of most of the modern parametric models.

The SIR model segments the population into three distinct compartments: susceptible (S), infected (I), and recovered (R). The transitions among these states are dictated by parameters reflecting the disease's and population's characteristics. The dynamics of the SIR model

are governed by the following differential equations:

$$\frac{dS}{dt} = -\beta \frac{SI}{N} \tag{6.1}$$

$$\frac{dI}{dt} = \beta \frac{SI}{N} - \gamma I \tag{6.2}$$

$$\frac{dR}{dt} = \gamma I \tag{6.3}$$

where:

- $S(t)$ is the number of susceptible individuals,
- $I(t)$ is the number of infectious individuals,
- $R(t)$ is the number of recovered individuals,
- N is the total population, $N = S + I + R$,
- β is the effective contact rate of the disease,
- γ is the recovery rate, representing the reciprocal of the infectious period.

The parameters β and γ can be estimated with maximum likelihood estimators, Bayesian regression, and other statistical learning algorithms. Additionally, researchers have proposed more complex statistical models to adapt to influenza forecasting, such as, Gaussian Process regression [Zimmer and Yaesoubi, 2020], Dynamic Bayesian models [Osthus et al., 2019], and the Dante system [Osthus and Moran, 2021].

6.3 Influenza Forecasting with Machine Learning

The second category comprises models based on machine learning techniques, which map input features to outputs using architectures that capture temporal and spatial dependencies. Recurrent neural networks have been effectively used for both global epidemic forecasting and more weekly influenza predictions in the United States [Amendolara et al., 2023, Wu et al., 2021]. Additionally, Yang et al. [2023] used recurrent neural networks on climate, demography, and search engine data for accurate prediction of influenza-like illness (ILI %) in China. Lastly, random forest algorithms have also been employed to predict influenza activity in the subtropical zones of Eastern China [Liu et al., 2019a].

6.4 Influenza Dataset

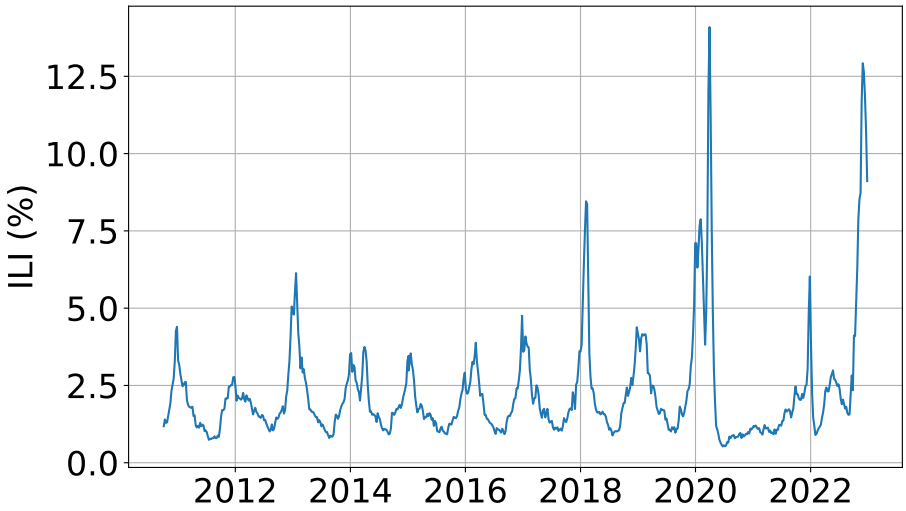


Figure 6.1: Influenza Like Illness (%) for New York City. The Influenza seasons show clear peaks during the winter until the end of 2019 and after that, the patterns became irregular due to COVID-19.

Another downstream predictive task we chose for the WeatherFormer was Influenza forecasting. The spread and severity of the influenza epidemic are strongly influenced by the mean temperature. Other weather variables such as humidity, precipitation, and wind speed also affect the process to varying degrees. Previously, several researchers, such as Amendolara et al. [2023] and Yang et al. [2023] have used weather as a data source for prediction. We decided to use only the mean temperature from NASA [2024] since our preliminary investigations on small models suggested that additional weather measurements introduce noise in the dataset.

6.4.1 Data Collection

We obtained our influenza dataset from Farrow et al. [2015], which includes weekly totals of Influenza-like Illness (ILI) cases and the percentage of ILI cases for New York City. The dataset spans 11 influenza seasons, from 2010-2011 to 2021-2022, and comprises weekly counts of patients, influenza cases, and the percentage of patients with ILI. We focused on predicting the percentage of ILIs among all patients. The dataset started on the week 40 of 2010. Due to a distribution shift attributed to COVID-19, data post-2020 were excluded.

For weather, we used our satellite-based weather data averaged over New York City. Amendolara et al. [2023] reported that temperature is a strong predictor of influenza cases followed by precipitation and wind speed. Average cooling and heating days are also important as these metrics relate to the time people spend indoors, a factor influencing transmission. We, however, discovered that only average temperature was sufficient for good prediction. Adding more measurements did not seem to improve the performance. This could be because our data source was from the satellite, which might be too imprecise for the task.

6.4.2 Train-Validation Split

We divided the remaining data into training and validation sets in four sequential phases: data from 2010 to 2015 was used for training and data from 2016 was used for validation. This pattern was repeated with training extended by one year each time, and validation occurring in the subsequent year. This resulted in four training and validation datasets. The best model's performance on the validation sets was averaged. Following the methodology of Amendolara et al. [2023] and others in the field, we reported Mean Absolute Error (MAE) as our evaluation metric.

We evaluated each model using a rolling forecasting approach. Models received weather data and historical ILI cases for a fixed number of past weeks and were then tasked with predicting ILI cases for the subsequent 10 weeks without access to future weather information. Following each prediction cycle, the input window was shifted forward by one week, generating a new 10-week forecast that extended the previous one. This process was repeated which yielded 52 prediction tasks per year, each spanning 10 weeks.

6.5 Finetuning WeatherFormer for Influenza Forecasting

Like the yield prediction problem, we process the weather data with the WeatherFormer and influenza data separately. First, weather data is processed through WeatherFormer to extract features. These features, along with historical ILI data and patient counts, are passed into another transformer model to predict the percentage of ILI cases.Figure 6.2 depicts a forward pass through the model. We standardized all input data before training. We found that the historical window of input data critically influenced model performance. Consequently, we optimized this hyperparameter by exploring values within the range of

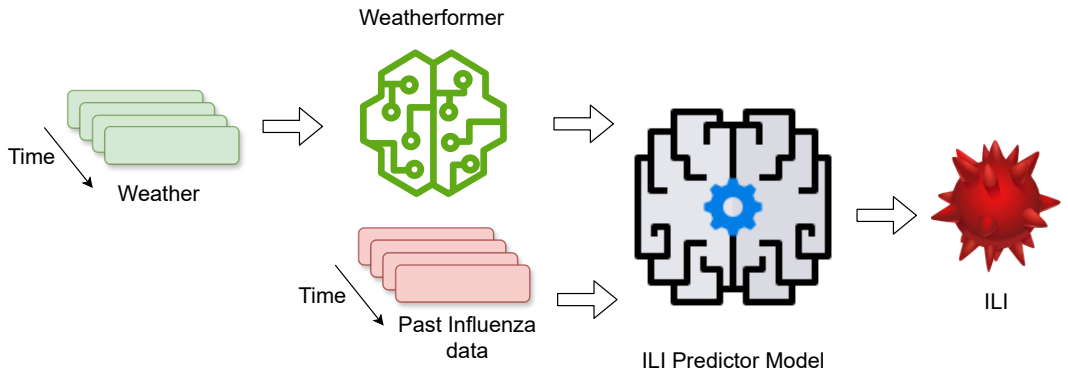


Figure 6.2: Influenza Like Illness (ILI) percent predictor architecture utilizing WeatherFormer. The weather measurements are first processed with WeatherFormer to extract useful features. These weather features and the past influenza data are processed by another transformer to predict ILI percent for the next 10 weeks.

$105, 110, 115, \ldots, 135$ weeks.

6.6 Baseline Models

For our baselines, we trained a least square linear regression model, an AutoRegressive Integrated Moving Average (ARIMA) [Harvey, 1990] model, and two transformer encoder models. Inspired by the autoregressive models, we added the most recent week's ILI value to the first predicted ILI value from the model. This helped to ground the predictions and led to considerably better performance.

Out of the two transformer models, one was trained only with the past ILI percentage and the total number of patients. The other transformer model utilized both the past influenza data and the weather data. Training two transformer models allows us to quantify the effects of weather on ILI percentage forecasting. For all of the baseline models, we experimented with the historical window length from 105 to 135 and reported the best performance.

6.7 Optimization

The final transformer model in each case had three layers and a hidden dimension of 64. These parameters were chosen after testing with hidden dimension sizes of 8, 16, 32, and 64, and 1-4 layers. Each model was tasked with forecasting the ILI percentage for a future period of 10 weeks. For the transformer-based models, Mean Squared Error (MSE) served

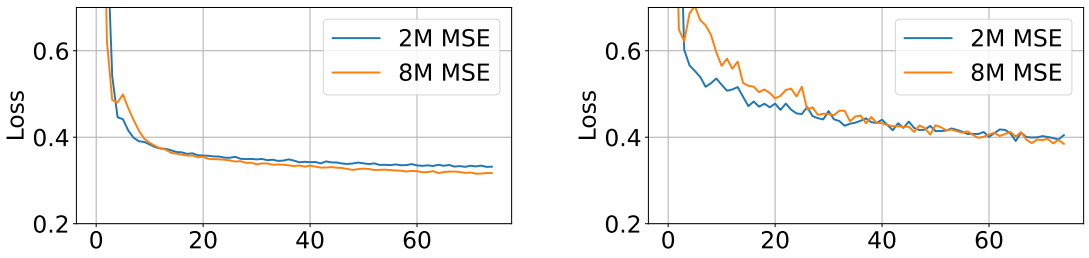
as the loss function. Optimization was carried out using the Adam optimizer with an initial learning rate of 0.0009 and for 25 epochs. We further employed an exponential learning rate scheduler that incorporated a 5-epoch warm-up phase followed by a decay factor of 0.95. Mean absolute error of 1, 5, and 10 weeks ahead predictions were reported in Table 7.2.

Chapter 7

Results and Discussion

7.1 Pretraining

We observe nearly identical performance on the validation set for both models during pretraining. This could be because the larger model requires more data to be effective. We



(a) Training. (b) Validation

Figure 7.1: Comparison between the 2M and 8M model losses during pretraining. Both models performed similarly in the validation dataset.

also observed that the input scaler weights shifted away from 1 during training. For the 2M parameter model, the mean value for the input scalers was 0.346, and the mean value for the weekly input scalers was 0.682, suggesting that the model learned to differentiate between data from different temporal granularities and also that the standard deviation of 1 for each input may not be ideal for the optimal weather representation.

Similarly, for the 8M parameter model, the mean of the daily input scalers was 0.437 and the mean of the weekly input scalers was 0.586.

It is important to note that in both cases the weekly input scaler was higher than the daily

input scaler, suggesting that the model recognized that the weekly input scalers need to have higher variance to account for the averaging over the week.

7.2 Yield Prediction

Table 7.1: Comparison of Validation RMSE for county-level soybeans yield forecasting. The WeatherFormer models are pretrained and the mean and the standard deviation of the dataset are 38.5 Bu/Acre and 11.03 Bu/Acre, respectively.

Model	**Validation RMSE**
Linear Regression	6.54
CNN-RNN	5.33
CNN-Transformer	5.13
WF-2M + Linear	5.15
WF-8M + Linear	5.27
WF-2M + Transformer	**4.83**
WF-8M + Transformer	4.84

In Table 7.1, we observe that adding a transformer with spatiotemporal encoding improves the performance of the CNN-Transformer model over the CNN-RNN model. However, its CNN module cannot extract the optimal representations from the weather inputs. On the other hand, WeatherFormer+Linear models can represent weather well, however, since they lack the transformer component to take advantage of the temporal trends in yield, they also show suboptimal performance. Only when the WeatherFormer model and the transformer model are combined in WeatherFormer+Transformer models, the performance is optimum.

We observed no difference between the best 2M and the best 8M parameter model performances. This could be because either the 8M parameter model requires more pretraining data for better performance, or the contribution of weather in yield prediction is already sufficiently captured with the 2M parameter model.

The performance of the WeatherFormer + Transformer models is remarkable because the WeatherFormer was pretrained using a different weather data source. However, the models still showed performance improvement after pretraining, suggesting that the WeatherFormer models learned good weather representations during pretraining and those representations were beneficial for the yield prediction task.

7.3 Influenza Forecasting

Table 7.2: Comparison of Validation MAE for ILI (%) forecasting. ARIMA models do very well for short-term prediction but their performance falls short in medium and long term predictions. Transformer models perform well in all three prediction tasks. The target variable had a mean of 2.43% and a standard deviation of 1.73.

Model	**+1 Week MAE**	**+5 Weeks MAE**	**+10 Weeks MAE**
Linear Regression	0.392	0.437	0.456
ARIMA	0.229	0.433	0.454
Transformer (without weather)	0.193	0.328	0.371
Transformer	0.189	0.294	0.320
WF-2M + Transformer	**0.186**	**0.277**	**0.297**
WF-8M + Transformer	0.188	0.291	0.329

We note that short-term influenza-like illness (ILI) behavior is predominantly influenced by recent influenza cases, as evidenced in Table Table 7.2. Here, we trained two identical transformer models, one with and one without weather data, and observed nearly identical performance in 1-week forecasts, indicating a minimal impact of weather data on such short-term predictions. This observation is further supported by the performance of the WeatherFormer model, which only slightly outperforms the baseline transformer model.

However, in the medium-term (5 weeks) and long-term (10 weeks) prediction, the weather becomes increasingly important. This is also reflected by the progressively worse performances of the Autoregressive Integrated Moving Average (ARIMA) model and the growing performance gap between the two baseline transformer models. We also observe that the pretrained WeatherFormer model performs progressively better, reflecting the growing importance of weather.

Chapter 8

Ablation Studies

In this chapter, we study the contribution of the various aspects WeatherFormer for its performance improvement.

8.1 Effects of Pretraining

We have trained WeatherFormer on a large pretraining dataset. It is important to question if this was even necessary or if the model would perform just as well without the lengthy pretraining process. For this, we train another set of WeatherFormer+Transformer models for both tasks, but this time the WeatherFormer component was not pretrained and the entire model was trained at once and only on the small, task-specific dataset.

Table 8.1: Comparison of Validation RMSE for county-level soybeans yield forecasting without pertaining. The large models overfit the training data and perform even worse than the least square linear regression on the validation dataset.

Model	**Best Validation RMSE**
Linear Regression	6.21
WeatherFormer-2M + Transformer	6.56
WeatherFormer-8M + Transformer	6.61

We observe in Table 8.1 that for the yield prediction problem, the large transformer models struggled to converge and their performances were even worse than the linear models. This highlights the problem of attempting to fit a large model in a small dataset, which often results in overfitting. In contrast, the previous pretrained WeatherFormer showed far better performance on the same dataset.

Table 8.2: Comparison of Validation MAE for ILI (%) forecasting. The WeatherFormer models are pretrained on the satellite-based weather measurements over the United States, Central America, and South America.

Model	**+1 Week MAE**	**+5 Weeks MAE**	**+10 Weeks MAE**
Linear Regression	0.392	0.437	0.456
WF-2M + Transformer	0.188	0.295	0.318
WF-8M + Transformer	0.197	0.375	0.436

On the influenza dataset, however, we observe that WeatherFormer2M+Transformer model converged and performed well. We believe this is because influenza cases exhibit far more predictable patterns. It is nonetheless noticeable that the performance of the untrained models was worse than the performance of the pretrained models for 5 weeks ahead and 10 weeks ahead prediction, suggesting improvement from the pretraining process. However, for the 1 week ahead predictions, the performance of the 2M parameter model was almost identical to the pretrained one, again suggesting a minimal impact of weather for short-term influenza forecasting.

8.2 Effects of the Architectural Innovation and Novel Pretraining Task

Next, we directly compare a finetuned WeatherFormer model's performance with a BERT-like transformer encoder model pretrained on the same dataset. We call this model WeatherBERT.

The WeatherBERT model lacks the feature mask, spatiotemporal encoding, and the input scalers that were introduced for WeatherFormer and it was trained with masked language modeling where 15% tokens were masked instead of the masked feature prediction for WeatherFormer. This model also has 2M parameters and has the same hidden size and the same number of layers as the WeatherFormer-2M.

Like the WeatherFormer models, we trained WeatherBERT for 75 epochs with Adam optimizer, 10 warmup epochs, and a learning rate decay factor of 0.99. We observed that the weather representations learned by the WeatherBERT were inferior to that of the WeatherFormer for downstream prediction tasks.

8.2.1 Yield Prediction

For the yield prediction task, we ran a similar experiment as before. The pretrained WeatherBERT-2M model was used in conjunction with the second transformer model for yield prediction. All other experimental parameters were unchanged. Consequently, the performance below on Table 8.3 compares the usefulness of the features learned by WeatherFormer and WeatherBERT.

Table 8.3: Comparison of Validation RMSE for county-level soybeans yield forecasting without pertaining. Representations learned by WeatherBERT showed inferior performance for yield prediction compared to WeatherFormer.

Model	**Best Validation RMSE**
WeatherFormer-2M + Transformer	4.83
WeatherBERT-2M + Transformer	5.22

We suspect that both masked language modeling and the lack of spatiotemporal encoding caused WeatherBERT to learn inferior features. Since the yield prediction dataset only contains six weather measurements, many features are completely masked during finetuing step. Since WeatherBERT did not observe this type of masking during the pretraining phase, it goes out of distribution. Secondly, without the spatiotemporal encoding and input scalers, WeatherBERT cannot properly use the location information and cannot distinguish between data from different temporal granularities. This causes further performance degradation.

8.2.2 Influenza Forecasting

Table 8.4: Comparison of Validation MAE for ILI (%) forecasting. Both models have the same hidden size, number of heads, and number of layers and they were pretrained for the same number of epochs on the same pretraining dataset.

Model	**+1 Week MAE**	**+5 Weeks MAE**	**+10 Weeks MAE**
WF-2M + Transformer	0.186	0.277	0.297
WeatherBERT-2M + Transformer	0.186	0.288	0.317

We can see that the WeatherFormer does progressively better at the longer-term predictions. As we have discussed before, the longer-term predictions have more influence from the weather, which again hints improved understanding of the weather by the WeatherFormer model compared to WeatherBERT.

Chapter 9

Conclusion and Future Work

9.1 Summary of Findings

- We have pretrained two WeatherFormer models of size 2M and 8M with a novel pretraining task, architectural innovations on a large satellite-based dataset.
- We have published the processed pretraining dataset, code, and the pretrained models.
- We have demonstrated its downstream performance for soybean yield prediction in the US corn belt and influenza forecasting in New York City.
- We have demonstrated that pretraining the model is necessary and leads to better downstream task performance.

9.2 Limitations

While the pretrained WeatherFormer has demonstrated improved performance in crop yield prediction and influenza forecasting, several limitations merit discussion:

1. **Geographical Coverage:** Our model training was confined to satellite data exclusively from the continental United States, Central America, and South America. Expanding the geographical scope to include other regions could enhance the model's applicability and robustness across different climatic zones globally. Furthermore, we relied solely on satellite data from the NASA Power Project [NASA, 2024], and more surface-level weather data sources could have been used in pretraining. Addressing this could improve the model's accuracy and generalization capabilities.

2. **Computational Resources and Optimization:** The model's training was constrained by available computational resources, limiting our ability to perform an exhaustive grid search for hyperparameter optimization. Instead, we adopted a sequential, greedy approach to optimize hyperparameters individually, which may not yield an optimal global configuration. This approach, while pragmatic, might compromise the model's overall performance potential.

3. **Data and Task Specificity:** Our experiments were limited to soybean yield predictions within the US corn belt and influenza forecasting in New York City, dictated by data availability. Testing the model across varied applications and more diverse geographic locations would provide a more comprehensive validation of its utility and scalability.

4. **Model Maintenance and Updating:** Given the dynamic nature of weather data, the model requires regular updates to maintain its relevance. This necessitates ongoing fine-tuning with new weather data, potentially once every year, to incorporate recent patterns and anomalies. Such continuous updating will require additional computational resources and logistical planning.

9.3 Future Work Directions

To address the above limitations, we propose the following research avenues for exploration:

1. **Expanding the Pretraining Dataset:** While our current model demonstrates promising results, we believe that incorporating a more extensive and diverse pretraining dataset can further enhance its performance. Sources of weather data could include additional weather measurements from NASA Power Project, CIMP-6 [Eyring et al., 2016], and ERA-5 [Hersbach et al., 2020].

2. **Diverse Applications:** In this study, we focused on soybean yield prediction and influenza forecasting. However, the potential applications of our model extend to a broader range of domains. Additional prediction tasks could be predicting yields for various crops, including corn, potatoes, rice, barley, and wheat. Furthermore, we envision the application of our approach to forecasting drought, coastal floods, and outbreaks of infectious diseases such as malaria and cholera on a global scale.

9.4 Code and Data Availability

The pretraining dataset is hosted on Hugging Face and the code is listed on GitHub.

9.5 Disclosures

ChatGPT [OpenAI, 2022] and Gemini [Anil et al., 2024] were used to write code during the research and to revise writing for this manuscript.

Appendix A

List of Weather Measurements in the Pretraining Data

Table A.1: Descriptions of the 31 Weather Variables with Their Units.

Parameter Name	**Symbol**	**Unit**
Temperature at 2 Meters	T2M	°C
Temperature at 2 Meters Maximum	T2M_MAX	°C
Temperature at 2 Meters Minimum	T2M_MIN	°C
Wind Direction at 2 Meters	WD2M	Degrees
Wind Speed at 2 Meters	WS2M	m/s
Surface Pressure	PS	kPa
Specific Humidity at 2 Meters	QV2M	g/Kg
Precipitation Corrected	PRECTOTCORR	mm/day
All Sky Surface Shortwave Downward Irradiance	ALLSKY_SFC_SW_DWN	$MJ/m^2/day$
Evapotranspiration Energy Flux	EVPTRNS	$MJ/m^2/day$
Profile Soil Moisture (0 to 1)	GWETPROF	0 to 1
Snow Depth	SNODP	cm
Dew/Frost Point at 2 Meters	T2MDEW	°C
Cloud Amount	CLOUD_AMT	0 to 1
Evaporation Land	EVLAND	$kg/m^2/s \times 10^6$
Wet Bulb Temperature at 2 Meters	T2MWET	°C
Land Snowcover Fraction	FRSNO	0 to 1

All Sky Surface Longwave Downward Irradiance	ALLSKY_SFC_LW_DWN	MJ/m^2/day
All Sky Surface PAR Total	ALLSKY_SFC_PAR_TOT	MJ/m^2/day
All Sky Surface Albedo	ALLSKY_SRF_ALB	0 to 1
Precipitable Water	PW	cm
Surface Roughness	Z0M	m
Surface Air Density	RHOA	kg/m^3
Relative Humidity at 2 Meters	RH2M	0 to 1
Cooling Degree Days Above 18.3 C	CDD18_3	days
Heating Degree Days Below 18.3 C	HDD18_3	days
Total Column Ozone	TO3	Dobson units
Aerosol Optical Depth 55	AOD_55	0 to 1
Reference Evapotranspiration	ET0	mm/day
Vapor Pressure	VAP	Pa
Vapor Pressure Deficit	VAD	Pa

References

NAMH Abdullah, NC Dom, SA Salleh, H Salim, and N Precha. The association between dengue case and climate: A systematic review and meta-analysis. *One Health*, 15:100452, Oct 2022. doi:10.1016/j.onehlt.2022.100452.

A.T.M.S. Ahamed, N.T. Mahmood, N. Hossain, M.T. Kabir, K. Das, F. Rahman, and R.M. Rahman. Applying data mining techniques to predict annual yield of major crops and recommend planting different crops in different districts in bangladesh. In *Proceedings of the 2015 IEEE/ACIS 16th International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing, SNPD 2015*. IEEE, 2015. doi:10.1109/SNPD.2015.7176185. URL https://doi.org/10.1109/SNPD.2015.7176185.

A. B. Amendolara, D. Sant, H. G. Rotstein, et al. Lstm-based recurrent neural network provides effective short term flu forecasting. *BMC Public Health*, 23:1788, 2023. doi:10.1186/s12889-023-16720-6. URL https://doi.org/10.1186/s12889-023-16720-6.

Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, Orhan Firat, James Molloy, Michael Isard, Paul R. Barham, Tom Hennigan, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, Ryan Doherty, Eli Collins, Clemens Meyer, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, Jack Krawczyk, Cosmo Du, Ed Chi, Heng-Tze Cheng, Eric Ni, Purvi Shah, Patrick Kane, Betty Chan, Manaal Faruqui, Aliaksei Severyn, Hanzhao Lin, YaGuang Li, Yong Cheng, Abe Ittycheriah, Mahdis Mahdieh, Mia Chen, Pei Sun, Dustin Tran, Sumit Bagri, Balaji Lakshminarayanan, Jeremiah Liu, Andras Orban, Fabian Güra, Hao Zhou, Xinying Song, Aurelien Boffy, Harish Ganapathy, Steven Zheng, HyunJeong Choe, Ágoston Weisz, Tao Zhu, Yifeng Lu, Siddharth Gopal, Jarrod Kahn, Maciej Kula, Jeff Pitman, Rushin Shah, Emanuel Taropa, Majd Al Merey, Martin Baeuml, Zhifeng Chen, Laurent El Shafey, Yujing Zhang, Olcan Sercinoglu, and George Tucker. Gemini: A family of highly capable multimodal models, 2024.

Bobby Azad, Reza Azad, Sania Eskandari, Afshin Bozorgpour, Amirhossein Kazerouni, Islem Rekik, and Dorit Merhof. Foundational models in medical imaging: A comprehensive survey and future vision, 2023.

D.F. Balting, A. AghaKouchak, G. Lohmann, et al. Northern hemisphere drought risk in a warming climate. *npj Climate and Atmospheric Science*, 4:61, 2021. doi:10.1038/s41612-021-00218-2.

Himanshu Batra, Narinder Singh Punn, Sanjay Kumar Sonbhadra, and Sonali Agarwal. *BERT-Based Sentiment Analysis: A Software Engineering Perspective*, page 138–148. Springer International Publishing, 2021. ISBN 9783030864729. doi:10.1007/978-3-030-86472-9_13. URL http://dx.doi.org/10.1007/978-3-030-86472-9_13.

Zhen Bi, Ningyu Zhang, Yida Xue, Yixin Ou, Daxiong Ji, Guozhou Zheng, and Huajun Chen. Oceangpt: A large language model for ocean science tasks, 2024.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020.

CDC. Weekly U.S. Influenza Surveillance Report. https://www.cdc.gov/flu/weekly/index.htm, 2024. [Accessed 08-05-2024].

Eirini Christaki, Panagiotis Dimitriou, Katerina Pantavou, and Georgios K. Nikolopoulos. The impact of climate change on cholera: A review on the global status and future challenges. *Atmosphere*, 11(5), 2020. ISSN 2073-4433. doi:10.3390/atmos11050449. URL https://www.mdpi.com/2073-4433/11/5/449.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. Electra: Pre-training text encoders as discriminators rather than generators, 2020.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021.

Joris P.C. Eekhout and Joris de Vente. Global impact of climate change on soil erosion and

potential for adaptation through soil conservation. *Earth-Science Reviews*, 226:103921, 2022. ISSN 0012-8252. doi:https://doi.org/10.1016/j.earscirev.2022.103921. URL https://www.sciencedirect.com/science/article/pii/S0012825222000058.

Veronika Eyring, Sandrine Bony, Gerald A. Meehl, Catherine A. Senior, Bjorn Stevens, Ronald J. Stouffer, and Karl E. Taylor. Overview of the coupled model intercomparison project phase 6 (cmip6) experimental design and organization. *Geoscientific Model Development*, 9(5):1937–1958, 2016.

B. Jalalzadeh Fard, S. A. Hasan, and J. E. Bell. Climedbert: A pre-trained language model for climate and health-related text, 2022.

David C. Farrow, Logan C. Brooks, Ryan J. Tibshirani, and Roni Rosenfeld. Delphi epidata api. GitHub repository, 2015. URL https://github.com/cmu-delphi/delphi-epidata.

Nigel Fernandez, Aritra Ghosh, Naiming Liu, Zichao Wang, Benoît Choffin, Richard Baraniuk, and Andrew Lan. Automated scoring for reading comprehension via in-context bert tuning, 2023.

A.C. Harvey. Arima models. In John Eatwell, Murray Milgate, and Peter Newman, editors, *Time Series and Statistics*, The New Palgrave. Palgrave Macmillan, London, 1990. doi:10.1007/978-1-349-20865-4_2. URL https://doi.org/10.1007/978-1-349-20865-4_2.

Hans Hersbach, Bill Bell, Paul Berrisford, Shoji Hirahara, András Horányi, Joaquín Muñoz-Sabater, Julien Nicolas, Carole Peubey, Raluca Radu, Dinand Schepers, et al. The era5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730):1999–2049, 2020.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

Moshe B Hoshen and Andrew P Morse. A weather-driven model of malaria transmission. *Malaria Journal*, 3:32, Sep 2004. doi:10.1186/1475-2875-3-32.

J.H. Jeong, J.P. Resop, N.D. Mueller, D.H. Fleisher, K. Yun, E.E. Butler, and S.H. Kim. Random forests for global and regional crop yield predictions. *PLoS ONE*, 11(6), 2016. doi:10.1371/journal.pone.0156571. URL https://doi.org/10.1371/journal.pone.0156571.

C. Jia, Q. Cao, Z. Wang, A. van den Dool, and M. Yue. Climate change affects the spread of typhoid pathogens. *Microbial Biotechnology*, 17(2):e14417, 2024. doi:10.1111/1751-7915.14417.

Sule Kahraman. Validation, calibration, and uncertainty quantification of the wofost crop simulation model. Master of engineering in electrical engineering and computer science,

Massachusetts Institute of Technology, Cambridge, MA, June 2021. URL https://hdl.handle.net/1721.1/139245. Accessed: 2022-01-14T14:59:05Z.

B.A Keating, P.S Carberry, G.L Hammer, M.E Probert, M.J Robertson, D Holzworth, N.I Huth, J.N.G Hargreaves, H Meinke, Z Hochman, G McLean, K Verburg, V Snow, J.P Dimes, M Silburn, E Wang, S Brown, K.L Bristow, S Asseng, S Chapman, R.L McCown, D.M Freebairn, and C.J Smith. An overview of apsim, a model designed for farming systems simulation. *European Journal of Agronomy*, 18(3):267–288, 2003. ISSN 1161-0301. doi:https://doi.org/10.1016/S1161-0301(02)00108-9. URL https://www.sciencedirect.com/science/article/pii/S1161030102001089. Modelling Cropping Systems: Science, Software and Applications.

W. O. Kermack and A. G. McKendrick. A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, pages 700–721, 1927.

Saeed Khaki, Lizhi Wang, and Sotirios V. Archontoulis. A cnn-rnn framework for crop yield prediction. *Frontiers in Plant Science*, 10, 2020. ISSN 1664-462X. doi:10.3389/fpls.2019.01750. URL https://www.frontiersin.org/journals/plant-science/articles/10.3389/fpls.2019.01750.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017.

E. Kirezci, I.R. Young, R. Ranasinghe, et al. Projections of global-scale extreme sea levels and resulting episodic coastal flooding over the 21st century. *Scientific Reports*, 10:11629, 2020. doi:10.1038/s41598-020-67736-6.

Remi Lam, Alvaro Sanchez-Gonzalez, Matthew Willson, Peter Wirnsberger, Meire Fortunato, Ferran Alet, Suman Ravuri, Timo Ewalds, Zach Eaton-Rosen, Weihua Hu, Alexander Merose, Stephan Hoyer, George Holland, Oriol Vinyals, Jacklynn Stott, Alexander Pritzel, Shakir Mohamed, and Peter Battaglia. Graphcast: Learning skillful medium-range global weather forecasting, 2023.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations, 2020.

Nguyen Quoc Khanh Le, Quang-Thai Ho, Van-Nui Nguyen, and Jung-Su Chang. BERT-Promoter: An improved sequence-based predictor of DNA promoter using BERT pre-trained model and SHAP feature selection. *Comput Biol Chem*, 99:107732, July 2022.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

W. Liu, Q. Dai, J. Bao, W. Shen, Y. Wu, Y. Shi, K. Xu, J. Hu, C. Bao, and X. Huo. Influenza activity prediction using meteorological factors in a warm temperate to subtropical transitional zone, eastern china. *Epidemiology and Infection*, 147:e325, 2019a. doi:10.1017/S0950268819002140. URL https://doi.org/10.1017/S0950268819002140. PMID: 31858924; PMCID: PMC7006024.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019b.

Xin Man, Chenghong Zhang, Jin Feng, Changyu Li, and Jie Shao. W-mae: Pre-trained weather model with masked autoencoder for multi-variable weather forecasting, 2023.

B A McFarland, N AlKhalifah, M Bohn, J Bubert, E S Buckler, I Ciampitti, J Edwards, D Ertl, J L Gage, C M Falcon, S Flint-Garcia, M A Gore, C Graham, C N Hirsch, J B Holland, E Hood, D Hooker, D Jarquin, S M Kaeppler, J Knoll, G Kruger, N Lauter, E C Lee, D C Lima, A Lorenz, J P Lynch, J McKay, N D Miller, S P Moose, S C Murray, R Nelson, C Poudyal, T Rocheford, O Rodriguez, M C Romay, J C Schnable, P S Schnable, B Scully, R Sekhon, K Silverstein, M Singh, M Smith, E P Spalding, N Springer, K Thelen, P Thomison, M Tuinstra, J Wallace, R Walls, D Wills, R J Wisser, W Xu, C T Yeh, and N de Leon. Maize genomes to fields (g2f): 2014-2017 field seasons: genotype, phenotype, climatic, soil, and inbred ear image datasets. *BMC Research Notes*, 13(1):71, 2020. doi:10.1186/s13104-020-4922-8.

NASA. NASA Power API, 2024. URL https://power.larc.nasa.gov/docs/referencing/.

Emeka Ndulue and Ramanathan Sri Ranjan. Performance of the fao penman-monteith equation under limiting conditions and fourteen reference evapotranspiration models in southern manitoba. *Theoretical and Applied Climatology*, 143(3):1285–1298, Feb 2021. ISSN 1434-4483. doi:10.1007/s00704-020-03505-9. URL https://doi.org/10.1007/s00704-020-03505-9.

Tung Nguyen, Johannes Brandstetter, Ashish Kapoor, Jayesh K. Gupta, and Aditya Grover. Climax: A foundation model for weather and climate, 2023.

OpenAI. Introducing ChatGPT, 2022. URL https://openai.com/blog/chatgpt.

D. Osthus and K.R. Moran. Multiscale influenza forecasting. *Nature Communications*, 12:2991, 2021. doi:10.1038/s41467-021-23234-5. URL https://doi.org/10.1038/s41467-021-23234-5.

D. Osthus, J. Gattiker, R. Priedhorsky, and S. Y. Del Valle. Dynamic bayesian influenza

forecasting in the united states with hierarchical discrepancy (with discussion). *Bayesian Analysis*, 14:261–312, 2019.

Jaideep Pathak, Shashank Subramanian, Peter Harrington, Sanjeev Raja, Ashesh Chattopadhyay, Morteza Mardani, Thorsten Kurth, David Hall, Zongyi Li, Kamyar Azizzadenesheli, Pedram Hassanzadeh, Karthik Kashinath, and Animashree Anandkumar. Fourcastnet: A global data-driven high-resolution weather model using adaptive fourier neural operators. *arXiv preprint arXiv:2202.11214*, 2022.

George Philander. El niño and la niña. *American Scientist*, 77(5):451–459, 1989. ISSN 00030996. URL http://www.jstor.org/stable/27855934.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training, 2018.

G. Ruß and R. Kruse. Regression models for spatial data: An example from precision agriculture. In *Proceedings of the Advances in Data Mining Applications and Theoretical Aspects*, pages 450–463, 2010. doi:10.1007/978-3-642-14400-4_35. URL https://doi.org/10.1007/978-3-642-14400-4_35.

G. Ruß, R. Kruse, M. Schneider, and P. Wagner. Data mining with neural networks for wheat yield prediction. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 5077 LNAI, pages 47–56, Berlin, Heidelberg, 2008. Springer Berlin Heidelberg. doi:10.1007/978-3-540-70720-2_4. URL https://doi.org/10.1007/978-3-540-70720-2_4.

Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The graph neural network model. *IEEE Transactions on Neural Networks*, 20(1): 61–80, 2009. doi:10.1109/TNN.2008.2005605.

O. Tetens. Über einige meteorologische begriffe (on some meteorological terms). *Z. Geophys.*, 6:297–309, 1930.

C.A. van Diepen, J. Wolf, H. van Keulen, and C. Rappoldt. Wofost: a simulation model of crop production. *Soil Use and Management*, 5(1):16–24, 1989. doi:https://doi.org/10.1111/j.1475-2743.1989.tb00755.x. URL https://bsssjournals.onlinelibrary.wiley.com/doi/abs/10.1111/j.1475-2743.1989.tb00755.x.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017.

Nicolas Webersinke, Mathias Kraus, Julia Anna Bingler, and Markus Leippold. Climatebert: A pretrained language model for climate-related text, 2022.

Dongxia Wu, Liyao Gao, Xinyue Xiong, Matteo Chinazzi, Alessandro Vespignani, Yi-An

Ma, and Rose Yu. Deepgleam: A hybrid mechanistic and deep learning model for covid-19 forecasting, 2021.

L. Yang, G. Li, J. Yang, T. Zhang, J. Du, T. Liu, X. Zhang, X. Han, W. Li, L. Ma, L. Feng, and W. Yang. Deep-learning model for influenza prediction from multisource heterogeneous data in a megacity: Model development and evaluation. *Journal of Medical Internet Research*, 25:e44238, Feb 2023. doi:10.2196/44238. URL https://doi.org/10.2196/44238. PMID: 36780207; PMCID: PMC9972203.

Chuang Zhao, Bing Liu, Liujun Xiao, Gerrit Hoogenboom, Kenneth J. Boote, Belay T. Kassie, Willingthon Pavan, Vakhtang Shelia, Kwang Soo Kim, Ixchel M Hernandez-Ochoa, Daniel Wallach, Cheryl H. Porter, Claudio O. Stockle, Yan Zhu, and Senthold Asseng. A simple crop model. *European Journal of Agronomy*, 104:97–106, 2019. ISSN 1161-0301. doi:https://doi.org/10.1016/j.eja.2019.01.009. URL https://www.sciencedirect.com/science/article/pii/S1161030118304234.

Jinhua Zhu, Yingce Xia, Lijun Wu, Di He, Tao Qin, Wengang Zhou, Houqiang Li, and Tie-Yan Liu. Incorporating bert into neural machine translation, 2020.

Christoph Zimmer and Reza Yaesoubi. Influenza forecasting framework based on Gaussian processes. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 11671–11679. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/zimmer20a.html.