Norwegian University of Life Sciences

Masters Thesis 2024 30 ECTS
Faculty of Science and Technology

A comparative study of soil temperature models, including machine learning models

Mats Hoem Olsen
Master of Science in Data Science

Forword

Jeg vil takke alle som var rundt meg under skrivingen, for uten dere så ville jeg ikke ha klart å fullføre.

Contents

List of Algorithms

List of Figures

List of Tables

Todo list

Acronyms

Symbols | K | L | M | N | R

Symbols

R^2

Explained Variance. 11

$\log(\kappa)$

Log Condition number. 11

K

Kilden

Norwegian Institute of Bioeconomy Research Kilden. 5

L

LMT

Norwegian Institute of Bioeconomy Research LandbruksMeteorologisk service. 5, 7, 8

LSTM

Long Short Term-Memory. 4

M

MAE

Mean Absolut Error. 11

MET

The Norwegian Meteorological Institute. 5, 7, 8

N

NMBU

The Norwegian University of Life Science. 17

R

RMSE

Root Mean Square Error. 11

Abstract

This study focuses on 3 models that have been used in the literature to predict soil temperatures. The depths chosen as targets are 10cm, and 20cm in 4 regions; Innlandet, Østfold, Vestfold, and Trøndelag. In each region there are 4 stations with self-draining or saturated soils.

Oppsummering

Denne studien ser på modeler som predikerer jordtemperaturer

Fill out when soon done

Keywords: Soil temperature, Machine learning, regression,hour, weather forecasting data

1 Introduction

In agriculture soil temperature is one of the important parameters to put into consideration when thinking about pest prevention, conservation, and yield prediction. The reasoning for this is that knowing the soil temperature is knowing climate change [1], water management [2], yield [3], nitrogen processes [4] in the soil, calculation of plant-growth [5], when seeds start to sprout [5], potential flooding and erosions[6], and predicting when insect eggs hatch that were laid last winter. Being able to predict the soil temperature into the future will be a huge advantage for farmers, civilians, and scientists.

If it's important, why don't institutions measure it everywhere? There are several reasons for this, but a common reason is that it's expensive to install new equipment on old weather stations. Sometimes the weather station do have the sensors in the fields reading soil temperature at given levels, but due to technical misadventures and unforeseen phenomenons there might be gaps or misreadings that need to be replaced with approximations or NULL values[1]. There are algorithms, models, and statistical tools to interpolate these missing values but they have their drawbacks. For instance approximation by global mean, which is a common method used in timeseries[7]. This method is preserved global statistics, however does not represent local changes. Further more for a good estimation of soil temperature it is useful to include exogenous[2] features.

There has been done research into heat conductivity in soil that has lead to differential equations[8], however these equations[8, 9] are computationally expensive and difficult to simulate, or calculate[4]. To add to the complexity the heat dynamics change depending on soil temperature as they change the physical stucture of the soil

Find source!, if not remove

There are also massive developments in the types of models, one of which are ASPER that combines logical statments[3] with deep learning models to achieve better or similar results to "non-logical" deep learning models, but on fewer samples[10]. A study has been preformed with the model which shows that the model can reduse the number of samples/observations needed by a factor of 1/1000[11]. After an intervju with the study resercher (Machot), although the model needs a strict ruleset it is possible to incorporate baysioan statistics to make the model more general for more applications by weakening the ruleset and imply that the rules given might not be 100% accurate and can be relaxed. Another cutting-edge method used is attention-awareness, same method used in ChatGPT, and other modern AI technologies for generative media. This method has been utilised to predict soil temperatures and soil moisture[1].

[1] These values are different from 0 as they represent "no data" and can't be used to do calculations.
[2] Variable that can affect the model, but is not not directly described by the model.
[3] Statements can be tought of as formulas, nature laws, knowledge about the solution

A beneficial model would be one using the fewest number of parameters as possible while returning results within acceptable tolerances. This study will consider models that can use only time and air temperature as those two features are the most common measurements measured at weather stations, since soil temperature is not necessarily calculated as stated earlier. A good metric in this study will be considered to be a combination of Root Mean Square Error and Explained Variance (see section 3.5).

The goal of this study is to find which type of model is worth to do further research on. The models selected is a small poll of models in the literature that are being used to predict soil temperature using metrological observations. The scope of metrological is limited to time (as either day of the year, or the hour of the year) as these are common observations in the literature and in Norwegian weather stations.

2 Theory

This section discusses the theory behind the models used in the study, there are a few base models (section 2.1, and 2.3) based on the interpretation of the papers hand picked for this study

2.1 Linear regression

The regression model will be for the sake of convenience be expressed as the following expression

$$\vec{F}(\mathbf{A})\vec{\beta} = \vec{y} + \vec{\varepsilon}$$

Where $\vec{F}$ is a vector function with following domain $\vec{F} : \mathbb{R}^{m \times n} \rightarrow \mathbb{R}^{m \times p}$ where $m, n, p \in \mathbb{N}$, $\mathbf{A}$ is the data in matrix form with dimensions $\mathbb{R}^{m \times n}$, $\vec{\beta}$ is the regression terms, $\vec{y}$ is the target (TJM10 or TJM20), and $\vec{\varepsilon}$ is the residual error.

The $\vec{F}$ is not important, just that your data is shaped by a function.

Remove

This basic model to express the linearity of the components to soil temperature. This will function as the base model for regression models.

2.2 Plauborg linear regression model with Fourier terms

An improvemnt over an time independent linear regression model would be a time dependent linear regression model that takes not only current time into account of the calculations but also previus messurments. It is current knowlage that soil temperatures depends on previous temperatures and metrological phenomenons. In the paper "Simple model for 10 cm soil temperature in different soils with short grass" the author chose to extend the features from only air temperature at current time to include also previus days of year and the air temperature from those days. This means the following F function that Plauborg used would be

$$\vec{F} := [air_t, air_{t-1}, air_{t-2}, air_{t-3}, \sin(\omega t), \cos(\omega t), \sin(2 * \omega t), \cos(2 * \omega t)]^T$$

Where air_t is the air temperature at time t expressed in day of the year (0-365), ω is the angular frequency in radians per hour or radians per day, depending on the time unit. The sine/cosine elements in the F function represent the variations through the day by fitting $\vec{\beta}$ to the yearly variation. To adapt the authors model to an hourly time unit would be to either

1. Extend the F function to include a larger ω coefficient to reflect hourly oscillations in conjunction with daily fluxiations

2. Refit the Fourier terms with a larger ω coefficient to make the oscillations more representative of daily temperature changes.

The larger coefficient could be expressed as $\pi/12$ while the smaller ω for daily values would be rescaled to $2\pi/365$.

The problem with this approsh would be Fouriers Sine-Cosine series approximation which would suggest that Plauborg's method could be subject to overfitting with addition of more terms. On the other hand it gives us a way to compute the coefficients α_i and γ_i for sine and cosine terms respectively, though it would be more numerically stable with a pseudo-inverse computation or a max log likelihood approach.

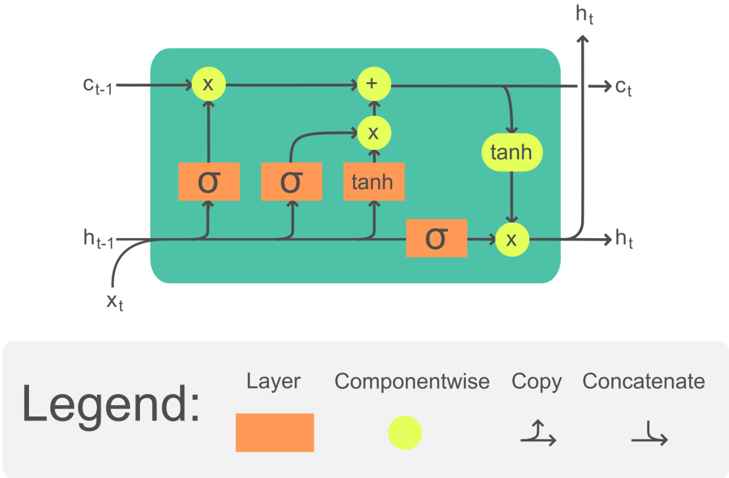


Figure 1: LSTM cell, Artist: Chevalier [13]

2.3 Long Short Term Memory model

Do proper credit of picture

When modeling soil temperature it is important to know the previus hours or days to predict the next timestep, for this a natural selection for a data driven model is a recurrent network. This type of network makes prediction based on previus timesteps in the data, however the longer timespan the model takse into account the less important are the erlier timesteps in the data.

To combat this there was developed an improved model called Long-Short Term Memory model[14] that deploys a memory cell that feeds information from earlier timesteps to the late ones. To make sure that redundant information or unimportant informantion don't get feed forward there are also forgetting gates that removes some of the newly learned patterns and integrates it into the memory cell.

Write more

The most common problem in Neural networks is the vanishing gradient problem where updating the first few layers of a large network becomes exponentially more difficult since the adjustments gets smaller and smaller for each layer towards the start rather than the reverse. Long Short Term-Memory changes this by caring information from the previous cells forward thereby allowing updating earlier cells with bigger impact than the standard approach[14]. LSTM is part of a family of Recurent Neural Network's that passes information to other cells in the same layer.

2.4 Attention aware LSTM model

Consider removing if not model

3 Method

3.1 Source of data

For this comparative study the following data sources will be used

1. Norwegian Institute of Bioeconomy Research LandbruksMeteorologisk service (LMT) 2. Norwegian Institute of Bioeconomy Research Kilden (Kilden) 3. The Norwegian Meteorological Institute (MET)

3.2 Dataset

The dataset is chosen from four regions in Norway; Innlandet, Vestfold, Trøndelag, and Østfold. From each region are four stations picked:

Region	Name	ID	Drain type	Soile category	Texture	MET name	Latitude	Longdetude
Innlandet	Apelsvoll	11	Selvdrenert	CM	17	SN11500	60,70024	10,86952
Innlandet	Fåvang	17	Selvdrenert	CM	15	SN13150	61,45822	10,1872
Innlandet	Ilseng	26	Selvdrenert	PH	17	SN12180	60,80264	11,20298
Innlandet	Kise	27	Vannmettet	GL	99	SN12550	60,77324	10,80569
Trøndelag	Kvithamar	57	Vannmettet	ST	16	SN69150	63,48795	10,87994
Trøndelag	Frosta	15	Selvdrenert	LP	13	SN69655	63,56502	10,69298
Trøndelag	Mære	34	Selvdrenert	RG	14	SN71320	63,94244	11,42527
Trøndelag	Rissa	39	Vannmettet	PL	13	SN71320	63,58569	9,97007
Vestfold	Lier	30	Vannmettet	ST	16	SN19940	59,79084	10,25962
Vestfold	Sande	42	Vannmettet	ST	16	SN26990	59,6162	10,22339
Vestfold	Tjølling	50	Selvdrenert	AR	13	SN27780	59,04641	10,12513
Vestfold	Ramnes	38	Vannmettet	ST	16	SN27315	59,38081	10,2397
Østfold	Rakkestad	37	Vannmettet	ST	18	SN3290	59,38824	11,39042
Østfold	Rygge	41	Selvdrenert	AR	13	SN17380	59,39805	10,75427
Østfold	Tomb	52	Vannmettet	ST	16	SN17050	59,31893	10,81449
Østfold	Øsaker	118	Vannmettet	ST	18	SN3370	59,31936	11,04221

Table 1: Station information from stations used in this study. The texture class is defined in this article: https://nibio.no/tema/jord/jordkartlegging/jordsmonnkart/dominerende-tekstur-i-overflatesjikt

All stations are sampled from the date[4] 03-01 to 10-31 from 2016 to 2020. The features rain (RR), mean soil temperature at 10cm (TJM10), mean soil temperature at 20cm (TJM20), and air temperature at 2m (TM) are sampled from the LMT database. The soil type, and soil texture is sampled from Kilden from Norwegian Institute of Bioeconomy Research.

3.2.1 Selection process

The selection process for finding these station can be compiled into these steps

1. Recommendation from Norwegian Institute of Bio-economy Research
2. Compute the missing values in the data
3. Missing values analyse

[4] Format month-day

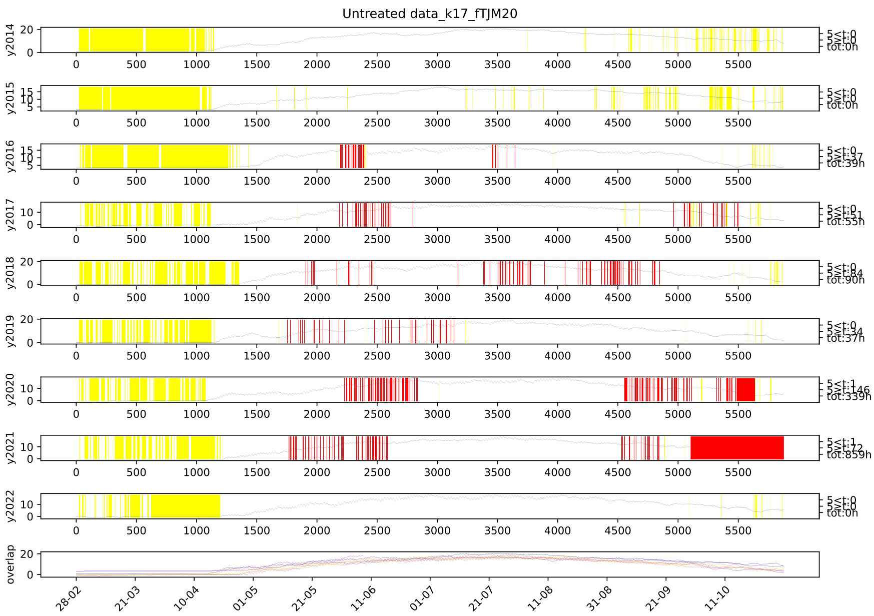


Figure 2: Visual representation of missing values at station 17 from 2014 to 2022 at the parameter "TJM20". The left numbers indicated how many hours that are missing and how many of them are shorter than or longer than 5 hours. The yellow markings indicate possible outliers based on the given year, all markings was checked if they were actual outliers. The red colouring indicate missing values in the data (represented in the data with code "NULL").

4. Searching LMT database for alternative station candidates if current data is insufficient

5. If some station was replaced the repeat step 2

The plots of stations follow a simple representation where the y-axis represent the year and the x-axis represent the index of the data as all tables are taken from the same period. A circle represent a singluar na values, while a band represent a series of 2 or more missing values. The colours represents the features used in this comperative study. This representation of the missing values will indicate sesonal, and systematic removal of data and give an overall indication of how much data is missing. To get further insight into the data a report is generated in parallel to the plots describing precise date and time of all values and which other parameter values is also missing values in the same period. See appendix **??** for the full detail of the report generation and appendix A for na-plots of the station chosen for this study.

3.2.2 Collection of data

The method used was a powershell[5] script that called the respective institutions servers using the "curl" program[6] to send an http request for the timeseries starting from 2014 to 2020 in the interval 1 of May to 31 of October. Code for data collection can be viewed in appendix **??**. The data is stores as an either a csv file or a json file for easy retrieval and manual control of values.

[5]Version 7.3.11

[6]curl 8.4.0 (Windows) libcurl/8.4.0 Schannel WinIDN

FROST	Description
Stations with rain	Requested rain data in millimeters.
Station ID	Sendt a request to LMT for station information using their remote API.
LMT	**Description**
Meteorological data	Requested soil temperature from 10cm depth, and 20cm depth and air temperature (2m), from 2014-03-01 to 2022-10-31.

Table 2: Description of what was requested from each server (FROST part of MET, LMT).

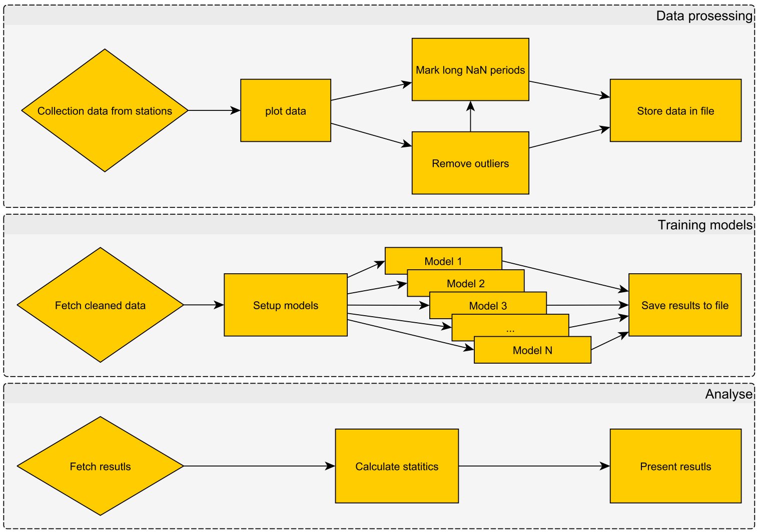


Figure 3: A surface level diagram of the methodology.

3.2.3 Labeling of stations between Nibio and MET

Since Nibio and MET have different names for the same stations one must compile a list that converts Nibio ID to MET ID. This was performed with these requests shown in table 2 where ID is the Nibio Id for the given station, Frost.ID is the MET id, ID. latitude is the latitude gathered from Nibio, ID.longitude is the longitude gathered from Nibio. These variables can be swaped out for the relevant station.

Move to caption

3.2.4 Storage of data

The storage of the data is done through two data structures; Hashmap and DataFrame from the package pandas. The transformation of data is done with a costume datatype called "DataFileHandler" which is converted to a module for convenience. The keys for the hashmap is chosen by the naming of the data files and the pattern given to the class. To escalete modeling the data will also be exported to a binary file for faster retrieval.

MORE DETAIL

Complete diagram

Technical overview of custom data structure The data structure used to store the data from the different stations is called "DataFileHandler" and stores the data in a tree-structure where indexes are dictated by the filename. It has several built-in functions to assist with data partitioning, and merging of data. This makes it easier to move and store all 846 720 observations from 16 station from all 4 regions[7]. Further more, the data structure has iteration functions so it is compatible with python's built in loops, and print functions.

3.3 Data cleaning and treatment

To use the data in this study it must be cleaned and treated for training. Though the data has been examined by the supplier, however it still had outliers that needed to be treated before modelling. For this reason several steps and methods is utilized in the prepossessing steps.

3.3.1 Outlier detection and removal

Revise so it reflects what you actually do

Though the data fetched from LMT is treated and controlled the external data from MET might not be, and this research project incorporated raw, untreated data from LMT to fill inn missing values.

The method to quickly find obvious outliers was to look at the following condition

$$\left|\frac{|\Delta T| - E(|\Delta T|)}{Var(|\Delta T|)}\right| > 4C^{\circ}$$

Make more clear

This condition looks at the absolute difference between consecutive measurements and calculates the z-score for each observation. It is expected that the change in temperature can't be too rapid. Further methods used to highlight potential outliers is

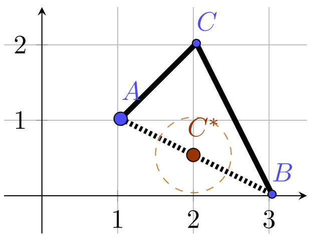


Figure 4: An simple outlier detection method utilizing a simple line to estimate where the expected point (C^*) is supposed to be. If observed point C falls outside the tolerance level (red dotted circle) then it is marked as an outlier.

3.3.2 Missing value imputation

The data has missing values, in particular during early Fall when there were sub-zero temperatures meaning any rain measurements done during this period would have unpredictable fluctuations since at negative temperatures water can freeze, get clogged up with residual biomaterial from the surrounding area . When interpolation the values the method chosen is a linear interpolation with a maximum period to interpoate is 5 hours for Soil temperatures, and 3 hours for air temperatures. The resoning for this is that the soil temperatures are more reliable

Rewrite this part to reflect what is going on

[7] there are 4 stations per region.

makeing it safer to interpolate without loosing too much information, while air temperaturses has a higher variance makeing it more difficult to interpolate without cutting values.

3.4 Setup of models

In general, write more on all subsections

The models are set up in according to the relevant paper the model is fetched from, alternatively reuse the code made by the author. When importing the data to the model there will be modifying to the original code to facilitate for the model as far as it goes. Any modifications will be in the appendix under section ??. For the convenience of the reader all code is using the sklearn estimator class to make all the models discuses in this study more user friendly and compatible with sklearns other functions. The details of the models will be discussed in section 2, this section discusses the setup and implementation of the models.[8]

3.4.1 Basic Linear model

The linear model (section 2.1) utilises in the study is created from the python model sklearn (or scikit-learn according to pythons package manager). The model is setup with standard parameters, and the data is fed into the model without scaling with fitted intercept coefficient.

3.4.2 Plauborg

The Plauborg regression will be formulated as a linear regression problem so that the LinearRegression function in the Sci-kit module can be used. For the parameters used in the paper[12] the F function defined in section 2.2 will be formulated with loops to give rise 3 more parameters for fine-tuning the model. Nan-values generated from the procedure get replaced with 0, since the data fed to the model is segnificantly larger than 10h (the minimum for the training is 24h).

3.4.3 LSTM

rewrite or possibaly move

Long Short-Term Memory (LSTM) networks are a specialized type of Recurrent Neural Network (RNN) designed to handle long-term dependencies in sequential data. Unlike regular RNNs, which often suffer from vanishing gradient problems, LSTMs utilize a more complex cell structure that allows them to capture long-term dependencies more effectively.

- **Input Gate (it)**: $a_i = \sigma(W_{xi} \cdot X_t + W_{hi} \cdot H_{t-1} + b_i)$
- **Forget Gate (f_t)**:
$$a_f = \sigma(W_{xf} \cdot X_t + W_{hf} \cdot H_{t-1} + b_f)$$
- **Cell State Update ($\tilde{C}_t$)**:
$$\tilde{C}_t = \tanh(W_{xc} \cdot X_t + W_{hc} \cdot H_{t-1} + b_c)$$
- **New Cell State (C_t)**:
$$C_t = f_t \cdot C_{t-1} + i_t \cdot \tilde{C}_t$$
- **Output Gate (o_t)**:
$$a_o = \sigma(W_{xo} \cdot X_t + W_{ho} \cdot H_{t-1} + b_o)$$

[8] Caution to the reader; The code used was run on the Linux subsystem (Debian) on windows due to the fact that the current version of tensorflow can't run on Windows.

- **Hidden State (H_t)**:

$$H_t = o_t \cdot \tanh(C_t)$$

Where:

X_t represents the input at time step (t).

H_{t-1} is the hidden state from the previous time step.

W and b are weight matrices and bias terms.

σ denotes the sigmoid activation function

tanh represents the hyperbolic tangent activation function.

LSTMs have proven effective in various tasks such as natural language processing, speech recognition, and time series prediction. They provide a powerful mechanism for modeling sequential data while mitigating the vanishing gradient problem commonly encountered in vanilla RNNs.

3.4.4 BiLSTM

rewrite

Bidirectional Long Short-Term Memory (BiLSTM) networks enhance traditional LSTMs by considering context from both forward and backward directions. Here's how they work:

BiLSTMs consist of two LSTM layers operating simultaneously:

- The forward LSTM processes the input sequence from the beginning to the end. - The backward LSTM processes the input sequence from the end to the beginning. By combining the outputs of these two LSTMs, BiLSTMs effectively capture information from both past and future context. - This context expansion is particularly useful in tasks where context matters in both directions:

Natural Language Processing (NLP): BiLSTMs excel in tasks like part-of-speech tagging, named entity recognition, and sentiment analysis. - Speech Recognition: Capturing context from both sides of an audio sequence improves accuracy. - Time Series Prediction: BiLSTMs enhance predictions by leveraging past and future data points. - Mathematically, the final hidden state of the BiLSTM at time step (t) is the concatenation of the forward and backward LSTM hidden states:

$$H_t = [H_t^{(f)}, H_t^{(b)}]$$

The resulting output can be used for downstream tasks such as classification, regression, or sequence labeling.

Information from earlier timesteps is important to say something about current timesteps, same can be said about the other time direction. To get the best of both direction one can use a bidirectional LSTM (BiLSTM) to combine the information from both approches. To make one of the models read the data backwards the model pipeline reverses the data and trains on that. A visualiation of this can be shown in the first layer in figure 5.

3.4.5 Modified BiLSTM

To investigate the posssibility of RNN a more complex model was developed to see if the introduction of more layers would improve the accuratsy of the base model.

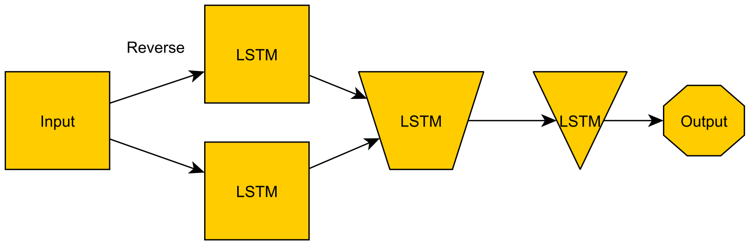


Figure 5: Simple diagram of the structure of the BiLSTM. The data from the input layer gets duplicated and passed to two LSTM's where one of them get the data in reverse. When both models have trained on the data they gets concatenated and passed to another LSTM that has half of the units as output. The final LSTM has a single cell as output leaving a single value as the prediction.

Complete caption, and improve figure

Soil temperatures depend on earlier timestamps, so to make accurate predictions, it's essential to include temperature data from $t, t-1, \ldots, t-k$. Evaluating the data both forwards and backwards helps identify features that are noticeable in specific directions. The BiLSTM model, implemented using TensorFlow's Keras module, consists of three layers:

EXPAND

- A Bidirectional LSTM layer for feature extraction.
- A regular LSTM layer to condense information with a tanh activation function
- An LSTM layer for prediction with a tanh activation function

The activation functions used are hyperbolic tangent and sigmoid. The output layer employs the identity function ($f(x) = x$). The hidden layer has half as many recurrent cells as the input layer, and the output layer condenses the hidden layer's output to a single value. For more details, refer to Figure 5 and Table 3a.

This model is an extention of "Modeling Hourly Soil Temperature Using Deep BiLSTM Neural Network"[5] written by Li *et al.*

3.5 Metrics

The metrics used in this study are

• Root Mean Square Error (RMSE) • Mean Absolut Error (MAE) • Explained Variance (R^2) • bias • Log Condition number ($\log(\kappa)$) • digit sensitivity

Soil temperatur as a differnet behavor than air temperature since energy (temperature) though the soil gets dampen and delayed. Since the data used in this study has outliers that was not cought during datatreatment, which has been addresed, the author of this study desided to include two more metrics that are not usually included in the evaluation; The log condition number, and digit sensitivity. Both metrics are based on the calculation of the condition number defined as

$$\kappa = \lim_{\varepsilon \to 0^+} \sup_{|\partial x| \leq \varepsilon} \frac{|f(x + \partial x) - f(x)|}{|f(x)|} * \frac{|x|}{|\partial x|} \tag{1}$$

Parameter name	Range
Epochs	[4,8]
Units	$\{2^6, 2^7, 2^8\}$
Lag time	$12i$ for $i \in [1, 14]$

(a) The search space for the modified BiLSTM model. The units define the number of LSTM cells used in the LSTM output. The lag time specifies how many hours the model will take into account when predicting soil temperature. The square brackets indicate an interval including endpoints, while the curly brackets indicate a list of elements.

Parameter name	Range
Sine terms	[1,10]
Cosine terms	[1,10]
Lag time	[1,14]

(b) The search space for the Plauborg model. The square brackets indicate an interval including endpoints. The "Lag time" indicated the number of time-steps before current time-steps do the model include $(t_{-1}, t_{-2}, \dots)$

Parameter name	Range
Epochs	[4,10]
Lead time	$\{24 * n \mid n \in [1, 7]\}$

(c) Parameter space for both LSTM and BiLSTM. The square brackets indicate an interval including endpoints. The "Lag time" indicated the number of time-steps before current time-steps do the model include $(t_{-1}, t_{-2}, \dots)$

Table 3: Parameter search space for the different deep learning models

This is not feasible to calculate since infinite calculations with infinitesimal numbers is not possible as per $May 5, 2024$ for simulation approach[9]. Therefore this paper uses algorithm 1 to approximate κ for all the models.

Algorithm 1: Method for calculating κ. $\mathcal{U}$ is a uniform random distrebution in a range.

Data: Data
Result: $\log(\kappa)$

Let κ_f be the function 1;
$\kappa \leftarrow 0$;
for $i \in 1 \dots |Data|$ **do**
  $\partial x \leftarrow \mathcal{U}_{[-\sqrt{\varepsilon/|Data|}, \sqrt{\varepsilon/|Data|}]}$;
  $k \leftarrow$ calculate with κ_f from x and $x + \partial x$;
  if $k > \kappa$ **then**
    $\kappa \leftarrow k$;
  end if
end for
return κ

The digit sensitivity is included to give an intitive understanding of κ and is computed simply as $\log_e(\kappa) + 1$. This number tells us the significant digit generated from the model. If the number is less than 0 then its the ith digit after the decimal point.

For the rest of the metrics, they are defined as follows

- RMSE $= \sqrt{\frac{\sum(y_{\text{pred}} - y_{\text{truth}})^2}{n}}$

[9] This calculation is possible for some models, for instance linear regression models when converted to the form $A\vec{\beta} = \vec{y}$.

- MAE $= \frac{\sum |y_{\text{pred}} - y_{\text{truth}}|}{n}$
- bias $= \frac{\sum (y_{\text{pred}} - y_{\text{truth}})}{n}$
- $R^2 = 1 - \frac{\sum (y_{\text{pred}} - y_{\text{truth}})^2}{\sum (y_{\text{pred}} - \vec{y})^2}$

Where $\vec{y}$ is the mean of the target, y_{pred} is the predicted data, and y_{truth} is the observed soil temperature.

3.5.1 Model training

The models get trained on air temperature, however the precise input for each model is not the same for all. The features used for each model are described in table 3 and their transformation in table 4.

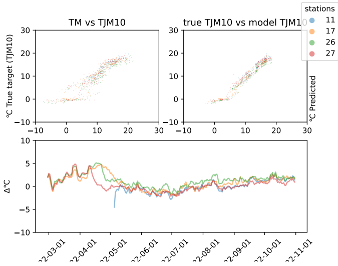
The models get a sample of the training data at the time due to the size and the amount for missing data (for example figure 2) The algorithm used to fetch reliable indexes are demonstrated at algorithm 2.

Algorithm 2: Find Non-NaN Ranges (Abstract)

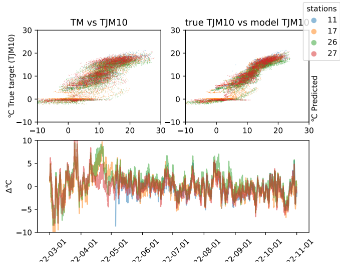
```
Input  : Input data data
Output: List of tuples: ranges

FindNonNaNRanges(data) ranges ← empty list;
start ← None;
for item in data do
    if itemisnotNan then
        if start is None then
            start ← item;
        end if
    end if
    else
        if start is not None then
            Add (start, item index - 1) to ranges;
            start ← None;
        end if
    end if
end for
if start is not None then
    Add (start, Last index) to ranges;
end if
return ranges;
```

model name	features	transformations
Linear regression	TM	Time get translated to hours by taking the day since new year and multiplying by 24 for then add the hour part.
Plauborg	Time, TM	Time get translated in two way; the current day since new year if looking at daily values, and hours since new year if looking at hourly predictions. When converting TM to daily values the hourly data get averaged.
BiLSTM	Time, TM	Time get translated to hours since new year.

Table 4: Parameters used for predicting soil temperatures at depth 10cm and 20cm.



(a) The daily model of Plauborg model



(b) The hourly model of Plauborg model

Figure 6: Comperation of daily versus hourly predictions

4 Results

Summerize results

Make sections

Include tables

4.1 Linear regression vs Plauborg

The global measure for the linear regression has an average error of $2.3C^{\circ} \pm 4.23C^{\circ}$ while the global messure of the Plauborg daily model has an average error of $0.6C^{\circ} \pm 1.96C^{\circ}$. Further more Plauborg has an hight R^2 value indicating that it follows the temperature changes in the soil better than just scaling the air temperature by a scaling factor.

4.2 Modification of Plauborg

The Plauborg model trained in Norway was found to only need 3 days (t_0, t_{-1}, t_{-2}) compared to [12] that needed 4 days $(t_0, t_{-1}, t_{-2}, t_{-3})$. However for the Fourier terms both models (Danish model and the Norwegian model) required 2 sine and cosine terms. For the 20cm target the models diverge in the sense of quantity of terms. It was found that the 20cm model needs 14 sine terms and 2 cosine terms, however only needs 2 days.

The modification to Plauborg's model is minor, by replacing the ω with a larger coefficient it can be used with hourly data. As seen in figure 6b the variation is stronger than 6a however the overall performance is comparable as seen in table 10 and table 8.

With modification to the model to accept hourly data it still preforms approximately as well as the daily data version. With a average error of $0.597C^{\circ} \pm 2.529C^{\circ}$ for TJM10 and $0.528C^{\circ} \pm 2.676C^{\circ}$ for TJM20. It was found that the modified Plauborg model only needs 2 sine terms to make a good prediction and 12h of air temperature which would translate to half a day instead of 3 days.

4.3 Deep learning models

4.4 Model comperison

The Plauborg has a clear advantage over linear regression.

model		52	37	50	38	57	34	27	17	average
lin-stat-10	R^2	0.221	0.473	0.584	0.388	0.295	-1.386	0.551	0.501	0.423
	MAE	2.889	3.503	2.702	3.741	3.356	3.415	3.236	3.432	3.267
	MSE	13.536	20.257	13.04	23.181	18.433	18.164	18.249	20.416	17.897
	bias	1.226	2.07	1.766	2.502	2.775	3.176	2.078	2.506	2.303
lin-stat-20	R^2	0.604	0.391	0.434	0.308	0.125	-2.248	0.415	0.336	0.308
	MAE	2.841	3.675	3.025	3.682	3.636	3.675	3.547	3.84	3.474
	MSE	12.648	22.605	16.383	23.345	21.664	21.007	21.832	25.492	20.286
	bias	0.559	2.174	2.207	2.601	3.153	3.471	2.535	2.939	2.487
Plauborg-day-stat-10	R^2	0.664	0.873	0.938	0.845	0.85	0.397	0.895	0.875	0.861
	MAE	1.837	1.755	1.105	1.908	1.547	1.687	1.627	1.78	1.621
	MSE	5.847	4.866	1.947	5.857	3.911	4.591	4.256	5.109	4.302
	bias	-0.636	0.373	-0.046	0.667	1.108	1.535	0.396	0.921	0.608
Plauborg-day-stat-20	R^2	0.83	0.905	0.946	0.912	0.863	0.316	0.901	0.844	0.876
	MAE	1.873	1.496	0.985	1.367	1.538	1.836	1.534	1.911	1.536
	MSE	5.427	3.523	1.566	2.96	3.391	4.425	3.703	5.995	3.647
	bias	-1.402	0.353	0.096	0.515	1.427	1.816	0.753	1.401	0.644
Plauborg-stat-10	R^2	0.636	0.857	0.884	0.754	0.677	-0.193	0.872	0.828	0.794
	MAE	1.964	1.804	1.472	2.368	2.263	2.306	1.724	1.979	1.926
	MSE	6.32	5.495	3.64	9.332	8.444	9.081	5.185	7.017	6.398
	bias	-0.349	0.237	0.558	1.363	0.677	0.845	0.163	0.065	0.597
Plauborg-stat-20	R^2	0.803	0.83	0.836	0.736	0.617	-0.547	0.839	0.762	0.756
	MAE	1.976	1.938	1.7	2.323	2.419	2.427	1.885	2.265	2.06
	MSE	6.272	6.317	4.736	8.9	9.482	10.006	6.027	9.136	7.16
	bias	-1.2	0.067	0.815	1.149	0.744	0.797	0.335	0.137	0.528
l1KerasBiLSTM-stats-10	R^2	-0.604	0.047	0.082	0.122	-0.038	-2.358	0.023	-0.088	-0.054
	MAE	4.444	5.04	4.619	4.984	4.268	4.112	5.278	5.651	4.768
	MSE	27.393	35.985	28.232	32.746	26.831	25.94	39.151	44.078	32.253
	bias	-0.853	0.429	-0.654	-0.05	1.435	2.278	0.582	1.674	0.528
l1KerasBiLSTM-stats-20	R^2	0.223	0.32	0.364	0.334	0.313	-1.231	0.346	0.305	0.284
	MAE	4.29	4.273	3.701	4.072	3.387	3.065	4.125	4.359	3.836
	MSE	24.585	25.046	18.206	22.271	16.933	14.596	24.221	26.55	20.854
	bias	-2.069	-0.305	-0.601	-0.189	0.784	1.274	0.118	0.83	-0.061
l2KerasBiLSTM-stats-10	R^2	-0.959	-0.115	-0.172	-0.095	-0.148	-2.523	-0.135	-0.181	-0.234
	MAE	4.869	5.431	5.149	5.613	4.534	4.273	5.648	5.861	5.143
	MSE	32.587	40.967	34.709	39.802	29.09	27.349	44.345	46.793	36.83
	bias	-1.547	-0.248	-1.436	-0.828	0.846	1.793	-0.126	1.04	-0.172
l2KerasBiLSTM-stats-20	R^2	-0.417	-0.112	-0.165	-0.096	-0.149	-3.111	-0.198	-0.248	-0.291
	MAE	5.52	5.466	4.967	5.195	4.537	4.447	5.698	6.036	5.219
	MSE	44.815	40.967	33.363	36.67	28.324	26.891	44.374	47.668	37.595
	bias	-2.054	0.067	-0.949	-0.465	1.49	2.397	0.528	1.692	0.195

model		41	118	42	30	39	15	26	11	average
lin-stat-10	R^2	0.593	0.306	0.506	0.588	0.213	-0.003	0.456	0.51	0.423
	MAE	3.07	3.486	3.525	3.106	3.193	3.141	3.651	2.713	3.267
	MSE	15.81	20.252	20.896	16.425	16.456	15.352	22.218	12.722	17.897
	bias	1.494	2.93	2.109	1.733	2.835	2.799	2.902	1.529	2.303
lin-stat-20	R^2	0.491	0.162	0.393	0.456	0.081	-0.241	0.302	0.35	0.308
	MAE	3.286	3.654	3.741	3.433	3.39	3.342	4.009	2.924	3.474
	MSE	18.048	22.338	23.645	19.934	18.574	17.627	26.732	14.598	20.286
	bias	1.677	3.208	2.364	2.015	3.083	3.006	3.282	1.692	2.487
Plauborg-day-stat-10	R^2	0.9	0.839	0.881	0.908	0.828	0.787	0.855	0.864	0.861
	MAE	1.587	1.697	1.75	1.519	1.455	1.428	1.937	1.504	1.621
	MSE	3.901	4.689	5.012	3.665	3.596	3.262	5.905	3.532	4.302
	bias	-0.293	1.137	0.333	-0.046	1.193	1.114	1.251	0.339	0.608
Plauborg-day-stat-20	R^2	0.914	0.886	0.901	0.91	0.852	0.801	0.833	0.866	0.876
	MAE	1.409	1.384	1.54	1.471	1.468	1.411	2.101	1.443	1.536
	MSE	3.056	3.034	3.866	3.3	2.99	2.827	6.393	3.011	3.647
	bias	-0.371	1.139	0.346	-0.014	1.402	1.215	1.578	0.463	0.644
Plauborg-stat-10	R^2	0.903	0.706	0.852	0.892	0.633	0.562	0.843	0.823	0.794
	MAE	1.519	2.322	1.885	1.555	2.151	1.991	1.918	1.666	1.926
	MSE	3.756	8.574	6.254	4.295	7.673	6.705	6.41	4.605	6.398
	bias	0.151	1.639	0.703	0.354	0.701	0.903	0.821	0.038	0.597
Plauborg-stat-20	R^2	0.872	0.656	0.807	0.859	0.615	0.484	0.801	0.755	0.756
	MAE	1.665	2.422	2.099	1.708	2.186	2.094	2.105	1.847	2.06
	MSE	4.559	9.173	7.5	5.181	7.784	7.321	7.603	5.504	7.16
	bias	0.13	1.722	0.746	0.428	0.633	0.827	0.892	-0.023	0.528
l1KerasBiLSTM-stats-10	R^2	0.042	0.055	0.069	0.039	-0.03	-0.516	-0.058	-0.243	-0.054
	MAE	5.201	4.322	5.356	5.259	3.734	3.973	5.452	4.846	4.768
	MSE	36.423	27.232	38.789	37.637	21.189	22.711	42.729	32.039	32.253
	bias	-0.744	0.66	-0.018	-0.414	1.523	1.198	1.636	-0.297	0.528
l1KerasBiLSTM-stats-20	R^2	0.317	0.333	0.33	0.331	0.339	0.058	0.317	0.155	0.284
	MAE	4.253	3.465	4.397	4.232	3.012	3.039	4.204	3.686	3.836
	MSE	23.978	17.674	25.895	24.313	13.251	13.248	26.035	18.914	20.854
	bias	-1.059	0.45	-0.342	-0.704	0.724	0.515	0.969	-0.949	-0.061
l2KerasBiLSTM-stats-10	R^2	-0.185	-0.14	-0.12	-0.17	-0.132	-0.696	-0.177	-0.431	-0.234
	MAE	5.702	4.805	5.834	5.721	3.929	4.159	5.742	5.172	5.143
	MSE	43.544	32.106	45.506	44.437	22.646	24.383	46.482	36.272	36.83
	bias	-1.529	-0.037	-0.808	-1.223	0.96	0.519	0.949	-1.03	-0.172
l2KerasBiLSTM-stats-20	R^2	-0.241	-0.215	-0.14	-0.26	-0.102	-0.867	-0.302	-0.753	-0.291
	MAE	5.645	4.903	5.719	5.812	3.984	4.442	6.108	5.394	5.219
	MSE	43.588	32.196	44.059	45.766	22.094	26.254	49.591	39.234	37.595
	bias	-1.264	0.157	-0.489	-0.893	1.599	1.087	1.52	-0.485	0.195

scope	spesific scope	RMSE °C	MAE°C	bias °C	log(κ(model))	digit sensitivity	R^2
global	—	2.529	1.926	0.597	-0.449	-1	0.794
region	Østfold	2.448	1.894	0.512	-0.448	-1	0.816
region	Vestfold	2.412	1.81	0.733	-0.445	-1	0.846
region	Trøndelag	2.822	2.176	0.781	-0.447	-1	0.547
region	Innlandet	2.382	1.805	0.312	-0.442	-1	0.847
local	52	2.514	1.964	-0.349	-0.439	-1	0.636
local	41	1.938	1.519	0.151	-0.446	-1	0.903
local	37	2.344	1.804	0.237	-0.45	-1	0.857
local	118	2.928	2.322	1.639	-0.446	-1	0.706
local	50	1.908	1.472	0.558	-0.441	-1	0.884
local	42	2.501	1.885	0.703	-0.441	-1	0.852
local	38	3.055	2.368	1.363	-0.441	-1	0.754
local	30	2.072	1.555	0.354	-0.439	-1	0.892
local	57	2.906	2.263	0.677	-0.444	-1	0.677
local	39	2.77	2.151	0.701	-0.439	-1	0.633
local	34	3.013	2.306	0.845	-0.447	-1	-0.193
local	15	2.589	1.991	0.903	-0.444	-1	0.562
local	27	2.277	1.724	0.163	-0.439	-1	0.872
local	26	2.532	1.918	0.821	-0.444	-1	0.843
local	17	2.649	1.979	0.065	-0.439	-1	0.828
local	11	2.146	1.666	0.038	-0.442	-1	0.823

Table 5: Hourly Plauborg model results.

scope	spesific scope	RMSE °C	MAE°C	bias °C	log(κ(model))	digit sensitivity	R^2
global	—	2.074	1.621	0.608	-1.261	-2	0.861
region	Østfold	2.168	1.704	0.24	-1.263	-2	0.856
region	Vestfold	2.022	1.564	0.219	-1.25	-2	0.892
region	Trøndelag	1.957	1.528	1.235	-1.27	-2	0.782
region	Innlandet	2.165	1.71	0.714	-1.264	-2	0.873
local	52	2.418	1.837	-0.636	-1.264	-2	0.664
local	41	1.975	1.587	-0.293	-1.259	-2	0.9
local	37	2.206	1.755	0.373	-1.261	-2	0.873
local	118	2.165	1.697	1.137	-1.259	-2	0.839
local	50	1.395	1.105	-0.046	-1.264	-2	0.938
local	42	2.239	1.75	0.333	-1.264	-2	0.881
local	38	2.42	1.908	0.667	-1.266	-2	0.845
local	30	1.914	1.519	-0.046	-1.255	-2	0.908
local	57	1.978	1.547	1.108	-1.266	-2	0.85
local	39	1.896	1.455	1.193	-1.267	-2	0.828
local	34	2.143	1.687	1.535	-1.252	-2	0.397
local	15	1.806	1.428	1.114	-1.262	-2	0.787
local	27	2.063	1.627	0.396	-1.266	-2	0.895
local	26	2.43	1.937	1.251	-1.264	-2	0.855
local	17	2.26	1.78	0.921	-1.268	-2	0.875
local	11	1.879	1.504	0.339	-1.266	-2	0.864

Table 6: Daily Plauborg model results.

5 Discussion

Write a discussion

5.1 The Autumn descrepensy

metion annomelies and resons

Show table of parameters

A phenomenon that arose during performance evaluations was that the linear models struggles with the Autumn season. The difference graphs showed a clear over or under estimation that are larger than 10σ. When investigating the coeffisents to the model this decrepsensy can be contributed to the intercept that during low temperature ($< 5c°$) giving either an over estimation or an under estimation. Further more when removing the calculation of the intersept the same phenomenon is still precent possibility due to the adaptation to summmer season.

5.1.1 Temperature seasons

In the diff plots there is a sutle distinction between Spring, Summer, and Autumn. This effect comes from the Winter where the snow lays ontop of the ground isolation the soil from air temperature changes creating a minimum temperature bound of around $0c°$.

5.2 Plauborg

The result of the modelling (table 9 to 12) show that modelig soil temperature without the inclusion of time is an inefficient, and inaccurate method of predicting soil temperatures.

In this study the original model, that was trained for daily values was converted to predict hourly data to see if the same formulation could be used to make predictions. When comparing the results shown in table 8 and table 7 to their daily counterpart it shows similar values showing that the model proposed in Plauborg can be extended to hourliy timeseries.

5.3 Explenation for good results of Plauborg

The inclusion of previous temperatures gives an improved estimation, even on hourly basis. The coefficents for both daily and hourly are observed to be <1 making it a mean temperature and the fourier terms would estimate the soil function(2)[19].

$$E_{\text{year}}(T) + e^{z/D} \sin(\omega t - z/D + \phi) \tag{2}$$

Since the term $exp(z/D)$ is constant we would be estimating $\sin(\omega t - Q) = \sin(\omega t)\cos(Q) - \sin(Q)\cos(\omega t)$, where Q is $z/D - \phi$ and is considered constant. This will be extrapolated to a simple sum of sines and cosines as the model does. Together the Plauborg model would estimate the approximation (4). The extra terms are nessesery to include external factors that affects the temperature (rain, soil type, atmosphere, etc).

$$E_{\text{year}}(T) + e^{z/D} \sin(\omega t - z/D + \phi) \tag{3}$$

$$\approx E_{\text{period}}(T) + \sum \alpha_i e^{z/D} \sin(-Q_i)\cos(i\omega t_i) + \sum \beta_j e^{z/D} \cos(-Q_j)\sin(i\omega t_j) \tag{4}$$

A possible reason for the residuals could be the soiltypes at the stations making so a universal, high accurasy model would not be feasible unless including other metrological metrics (air pressure, humidity, soil type, soil texture, etc...) or including other non-linear features ($\sqrt{\text{temperature}}$,ratio between temperature change in depth and time, etc...).

5.3.1 RNN results compared to other studies

The BiLSTM is an improvement over LSTM and the modified BiLSTM with layers is a clear indication that added complexity to a deep learning model is the way to go. This progression of improvements has been shown in other studies[1, 15–18]. None of the deep learning models has been optimised, however in according to earlier studies that focused on these types of model the authors did find that adding layers to the models does improve the model performance.

5.3.2 Further developments

It is a know fact, since 1849[10], that to know previous weather patterns will greatly improve prediction accuracy. To improve the accurasy even more a model can focus one specific patterns that has a big impact on the prediction. The ILSTM attempts to do this by including a new technique in data science called attention[20] that takes a collection of data and gives each element a weight associated with importance. When that paper was published it was focused on translation between English and German, however the paper published by Li *et al.* uses this novel technique to do both time and feature importance and from that make a prediction.

Make a comperason to BiLSTM

5.4 Future work

The models chosen in this study is not a representative sample of current knowledge of soil temperature modelling, and this study did not aim for optimizing the models beyond what the original authors have already done with the exception for base models used for comparison pouposus. A more comprehensive is needed of more complex models that utelises cutting edge technologies, techniques, and theory. One of which is logic based models, for instance ASPER[10] that tries to incorporate logical descriptions of the problem and limits the model for better or equal results based on fewer samples[11]. Another approch is to use the newest deep learning method of the attention mechanism[20] combined with recurrent neural networks to elivate the accuracy and speed of the model. As the author of the paper [1] has show great promise with that approch

[10] First weather prediction made by Joseph Henry in 1849

6 Conclusion

Soil temperature significantly impacts agriculture, influencing pest prevention, conservation, yield prediction, and more. Despite its importance, widespread measurement remains challenging due to cost limitations and technical issues. Interpolating missing data using methods like global mean approximation is common but has drawbacks including requiring a previus messurements of the soil temperature. Incorporating exogenous features can improve soil temperature estimation. Advancements in prediction and measurement are crucial for sustainable agriculture and accurate climate models.

Methods used in this study are

The prediction of soil temperatures is an important subject of study for the civil population, and the scientific community. For that reason a good model that can be generalised and applied globally is an desired thing.

The resutls show that good results can be achived with few parameters, however furter studies need to be done to see the effekt of adding more parameters. As for modeling; Adding time to a regression model does improve the model predictive power over a time independent model.

There is a clear advantage to further investigation into deep learning models as the models shows good results, as is show in other studies[1, 5, 15].

6.1 Limitations

This study faced a multitude of technical difficulties including,

- Getting the models to run
- Finding proper parameters
- Insufficient computing power
- Getting TensorFlow to work

7 Acknowledgements

7.1 Use of Artificial Intelligence in this paper

In this paper there has been used Artificial Intelligence (AI), specifically Bing Chat / Copilot hosted by Microsoft Cooperation with special agreement with The Norwegian University of Life Science (NMBU), for the following purposes:

1. Formalising sentences and rephrasing sentences.
2. Spellchecking
3. Code generation of basic consepts and structures (tree traversal, template for generic classes)

It is important to emphasize that our engagement with AI have been actively curated and verified with known information. All code underwent rigorous manual inspection within a dedicated testing environment. Furthermore, no confidential or sensitive information was shared with the AI; our interactions focused solely on broad topics and general inquiries. To validate the accuracy of AI-generated responses, we cross-referenced them with established research papers and textbooks.

8 Bibliography

References

[1] Qingliang Li, Yuheng Zhu, Wei Shangguan, Xuezhi Wang, Lu Li, and Fanhua Yu, "An attention-aware LSTM model for soil moisture and soil temperature prediction," *Geoderma*, volume 409, page 115 651, Mar. 2022, ISSN: 00167061. DOI: 10.1016/j.geoderma.2021.115651. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S001670612100731X (visited on 10/05/2023).

[2] Meysam Alizamir, Ozgur Kisi, Ali Najah Ahmed, Cihan Mert, Chow Ming Fai, Sungwon Kim, Nam Won Kim, and Ahmed El-Shafie, "Advanced machine learning model for better prediction accuracy of soil temperature at different depths," *PLOS ONE*, volume 15, number 4, Lei Lin, Ed., page 25, Apr. 14, 2020, ISSN: 1932-6203. DOI: 10.1371/journal.pone.0231055. [Online]. Available: https://dx.plos.org/10.1371/journal.pone.0231055 (visited on 09/29/2023).

[3] Ha Seon Sim, Dong Sub Kim, Min Gyu Ahn, Su Ran Ahn, and Sung Kyeom Kim, "Prediction of strawberry growth and fruit yield based on environmental and growth data in a greenhouse for soil cultivation with applied autonomous facilities," *Korean Journal of Horticultural Science and Technology*, volume 38, number 6, pages 840–849, Dec. 31, 2020, ISSN: 1226-8763, 2465-8588. DOI: 10.7235/HORT.20200076. [Online]. Available: https://www.hst-j.org/articles/doi/10.7235/HORT.20200076 (visited on 10/05/2023).

[4] Katri Rankinen, Tuomo Karvonen, and D. Butterfield, "A simple model for predicting soil temperature in snow-covered and seasonally frozen soil: Model description and testing," *Hydrology and Earth System Sciences*, volume 8, number 4, pages 706–716, Aug. 31, 2004, ISSN: 1607-7938. DOI: 10.5194/hess-8-706-2004. [Online]. Available: https://hess.copernicus.org/articles/8/706/2004/ (visited on 03/17/2023).

[5] Cong Li, Yaonan Zhang, and Xupeng Ren, "Modeling hourly soil temperature using deep BiLSTM neural network," *Algorithms*, volume 13, number 7, page 173, Jul. 17, 2020, ISSN: 1999-4893. DOI: 10.3390/a13070173. [Online]. Available: https://www.mdpi.com/1999-4893/13/7/173 (visited on 03/17/2023).

[6] Joris C. Stuurop, Sjoerd E.A.T.M. Van Der Zee, and Helen Kristine French, "The influence of soil texture and environmental conditions on frozen soil infiltration: A numerical investigation," *Cold Regions Science and Technology*, volume 194, page 103 456, Feb. 2022, ISSN: 0165232X. DOI: 10.1016/j.coldregions.2021.103456. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0165232X21002378 (visited on 01/31/2024).

[7] Mathieu Lepot, Jean-Baptiste Aubin, and François Clemens, "Interpolation in time series: An introductive overview of existing methods, their performance criteria and uncertainty assessment," *Water*, volume 9, number 10, page 796, Oct. 17, 2017, ISSN: 2073-4441. DOI: 10.3390/w9100796. [Online]. Available: http://www.mdpi.com/2073-4441/9/10/796 (visited on 02/17/2024).

[8] Tuomo Karvonen, *A model for predicting the effect of drainage on soil moisture, soil temperature and crop yield*. Otaniemi, Finland: Helsinki University of Technology, Laboratory of Hydrology and Water Resources Engineering, 1988, xvi, 215, Open Library ID: OL15197205M.

[9] Jean Baptiste Joseph Fourier and Alexander Freeman, *The analytical theory of heat*. New York: Cambridge University Press, 2009, OCLC: 880311398, ISBN: 978-1-108-00178-6.

[10] Trung Hoang Le, Huiping Cao, and Tran Cao Son, *ASPER: Answer set programming enhanced neural network models for joint entity-relation extraction*, version: 1, May 24, 2023. arXiv: 2305.15374[cs]. [Online]. Available: http://arxiv.org/abs/2305.15374 (visited on 03/12/2024).

[11] Fadi Al Machot, *Bridging logic and learning: A neural-symbolic approach for enhanced reasoning in neural models (ASPER)*, Dec. 18, 2023. arXiv: 2312.11651[cs]. [Online]. Available: http://arxiv.org/abs/2312.11651 (visited on 03/12/2024).

[12] Finn Plauborg, "Simple model for 10 cm soil temperature in different soils with short grass," *European Journal of Agronomy*, volume 17, number 3, pages 173–179, Oct. 2002, ISSN: 11610301. DOI: 10.1016/S1161-0301(02)00006-0. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1161030102000060 (visited on 03/17/2023).

[13] Guillaume Chevalier, *English: Schematic of the long-short term memory cell, a component of recurrent neural networks*, May 16, 2018. [Online]. Available: https://commons.wikimedia.org/wiki/File:LSTM_Cell.svg (visited on 12/10/2023).

[14] Sepp Hochreiter and Jürgen Schmidhuber, "Long short-term memory," *Neural Computation*, volume 9, number 8, pages 1735–1780, Nov. 1, 1997, ISSN: 0899-7667, 1530-888X. DOI: 10.1162/neco.1997.9.8.1735. [Online]. Available: https://direct.mit.edu/neco/article/9/8/1735-1780/6109 (visited on 10/18/2023).

[15] Yu Feng, Ningbo Cui, Weiping Hao, Lili Gao, and Daozhi Gong, "Estimation of soil temperature from meteorological data using different machine learning models," *Geoderma*, volume 338, pages 67–77, Mar. 2019, ISSN: 00167061. DOI: 10.1016/j.geoderma.2018.11.044. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0016706118311200 (visited on 03/17/2023).

[16] Hatice Citakoglu, "Comparison of artificial intelligence techniques for prediction of soil temperatures in turkey," *Theoretical and Applied Climatology*, volume 130, number 1, pages 545–556, Oct. 2017, ISSN: 0177-798X, 1434-4483. DOI: 10.1007/s00704-016-1914-7. [Online]. Available: http://link.springer.com/10.1007/s00704-016-1914-7 (visited on 09/28/2023).

[17] Qingliang Li, Huibowen Hao, Yang Zhao, Qingtian Geng, Guangjie Liu, Yu Zhang, and Fanhua Yu, "GANs-LSTM model for soil temperature estimation from meteorological: A new approach," *IEEE Access*, volume 8, pages 59 427–59 443, 2020, ISSN: 2169-3536. DOI: 10.1109/ACCESS.2020.2982996. [Online]. Available: https://ieeexplore.ieee.org/document/9045947/ (visited on 12/05/2023).

[18] Xuezhi Wang, Wenhui Li, Qingliang Li, and Xiaoning Li, "Modeling soil temperature for different days using novel quadruplet loss-guided LSTM," *Computational Intelligence and Neuroscience*, volume 2022, Xin Ning, Ed., pages 1–17, Feb. 17, 2022, ISSN: 1687-5273, 1687-5265. DOI: 10.1155/2022/9016823. [Online]. Available: https://www.hindawi.com/journals/cin/2022/9016823/ (visited on 12/05/2023).

[19] T. R. H. Holmes, M. Owe, R. A. M. De Jeu, and H. Kooi, "Estimating the soil temperature profile from a single depth observation: A simple empirical heatflow solution," *Water Resources Research*, volume 44, number 2, 2007WR005994, Feb. 2008, ISSN: 0043-1397, 1944-7973. DOI: 10.1029/2007WR005994. [Online]. Available: https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2007WR005994 (visited on 04/29/2024).

[20] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin, *Attention is all you need*, 2017. arXiv: 1706.03762[cs]. [Online]. Available: http://arxiv.org/abs/1706.03762 (visited on 04/16/2024).

Glossary

D | H | L | R

D

DataFrame
A table of values. The name is from the python library Pandas used in this study.. 7

H

Hashmap
A list of items where their unique placmnt in the list is detemend by their unique refrence key using a function that maps the key to a placement in the list.. 7

L

Long Short Term-Memory
A Recurent Neural Network with a memory cell to distribute information along the other RNN cells.. 4

R

Recurent Neural Network
A Neural network that passes information between cells in the same layers.. 4, D

Acronyms

Symbols | K | L | M | N | R

Symbols

R^2
Explained Variance. 11

$\log(\kappa)$
Log Condition number. 11

K

Kilden
Norwegian Institute of Bioeconomy Research Kilden. 5

L

LMT
Norwegian Institute of Bioeconomy Research LandbruksMeteorologisk service. 5, 7, 8

LSTM

Long Short Term-Memory. 4

M

MAE

Mean Absolut Error. 11

MET

The Norwegian Meteorological Institute. 5, 7, 8

N

NMBU

The Norwegian University of Life Science. 17

R

RMSE

Root Mean Square Error. 11

A Plots

B Tables

scope	spesific scope	RMSE °C	MAE°C	bias °C	log(κ(model))	digit sensitivity	R^2
global	—	2.676	2.06	0.528	-0.328	-1	0.756
region	Østfold	2.564	2	0.176	-0.324	-1	0.8
region	Vestfold	2.565	1.958	0.785	-0.317	-1	0.81
region	Trøndelag	2.938	2.279	0.75	-0.321	-1	0.477
region	Innlandet	2.612	1.997	0.379	-0.32	-1	0.799
local	52	2.504	1.976	-1.2	-0.327	-1	0.803
local	41	2.135	1.665	0.13	-0.322	-1	0.872
local	37	2.513	1.938	0.067	-0.322	-1	0.83
local	118	3.029	2.422	1.722	-0.323	-1	0.656
local	50	2.176	1.7	0.815	-0.327	-1	0.836
local	42	2.739	2.099	0.746	-0.323	-1	0.807
local	38	2.983	2.323	1.149	-0.333	-1	0.736
local	30	2.276	1.708	0.428	-0.328	-1	0.859
local	57	3.079	2.419	0.744	-0.33	-1	0.617
local	39	2.79	2.186	0.633	-0.322	-1	0.615
local	34	3.163	2.427	0.797	-0.329	-1	-0.547
local	15	2.706	2.094	0.827	-0.329	-1	0.484
local	27	2.455	1.885	0.335	-0.328	-1	0.839
local	26	2.757	2.105	0.892	-0.324	-1	0.801
local	17	3.023	2.265	0.137	-0.33	-1	0.762
local	11	2.346	1.847	-0.023	-0.33	-1	0.755

Table 7: Results from hourly version of the Plauborg model for 20cm depth.

scope	spesific scope	RMSE °C	MAE°C	bias °C	log(κ(model))	digit sensitivity	R^2
global	—	2.529	1.926	0.597	-0.449	-1	0.794
region	Østfold	2.448	1.894	0.512	-0.45	-1	0.816
region	Vestfold	2.412	1.81	0.733	-0.43	-1	0.846
region	Trøndelag	2.822	2.176	0.781	-0.439	-1	0.547
region	Innlandet	2.382	1.805	0.312	-0.448	-1	0.847
local	52	2.514	1.964	-0.349	-0.446	-1	0.636
local	41	1.938	1.519	0.151	-0.445	-1	0.903
local	37	2.344	1.804	0.237	-0.446	-1	0.857
local	118	2.928	2.322	1.639	-0.442	-1	0.706
local	50	1.908	1.472	0.558	-0.442	-1	0.884
local	42	2.501	1.885	0.703	-0.447	-1	0.852
local	38	3.055	2.368	1.363	-0.447	-1	0.754
local	30	2.072	1.555	0.354	-0.441	-1	0.892
local	57	2.906	2.263	0.677	-0.443	-1	0.677
local	39	2.77	2.151	0.701	-0.44	-1	0.633
local	34	3.013	2.306	0.845	-0.446	-1	-0.193
local	15	2.589	1.991	0.903	-0.444	-1	0.562
local	27	2.277	1.724	0.163	-0.444	-1	0.872
local	26	2.532	1.918	0.821	-0.442	-1	0.843
local	17	2.649	1.979	0.065	-0.44	-1	0.828
local	11	2.146	1.666	0.038	-0.443	-1	0.823

Table 8: Results from hourly version of the Plauborg model for 10cm depth.

scope	spesific scope	RMSE °C	MAE°C	bias °C	log(κ(model))	digit sensitivity	R^2
global	—	1.91	1.536	0.644	-1.918	-2	0.876
region	Østfold	1.94	1.541	-0.073	-1.917	-2	0.885
region	Vestfold	1.71	1.341	0.236	-1.899	-2	0.915
region	Trøndelag	1.843	1.56	1.461	-1.913	-2	0.794
region	Innlandet	2.16	1.735	1.02	-1.909	-2	0.863
local	52	2.33	1.873	-1.402	-1.908	-2	0.83
local	41	1.748	1.409	-0.371	-1.902	-2	0.914
local	37	1.877	1.496	0.353	-1.913	-2	0.905
local	118	1.742	1.384	1.139	-1.917	-2	0.886
local	50	1.251	0.985	0.096	-1.912	-2	0.946
local	42	1.966	1.54	0.346	-1.917	-2	0.901
local	38	1.721	1.367	0.515	-1.906	-2	0.912
local	30	1.817	1.471	-0.014	-1.918	-2	0.91
local	57	1.841	1.538	1.427	-1.919	-2	0.863
local	39	1.729	1.468	1.402	-1.912	-2	0.852
local	34	2.104	1.836	1.816	-1.91	-2	0.316
local	15	1.681	1.411	1.215	-1.903	-2	0.801
local	27	1.924	1.534	0.753	-1.905	-2	0.901
local	26	2.528	2.101	1.578	-1.912	-2	0.833
local	17	2.448	1.911	1.401	-1.912	-2	0.844
local	11	1.735	1.443	0.463	-1.91	-2	0.866

Table 9: Results from daily version of the Plauborg model for 20cm depth.

scope	spesific scope	RMSE °C	MAE°C	bias °C	log(κ(model))	digit sensitivity	R^2
global	—	2.074	1.621	0.608	-1.261	-2	0.861
region	Østfold	2.168	1.704	0.24	-1.27	-2	0.856
region	Vestfold	2.022	1.564	0.219	-1.263	-2	0.892
region	Trøndelag	1.957	1.528	1.235	-1.257	-2	0.782
region	Innlandet	2.165	1.71	0.714	-1.269	-2	0.873
local	52	2.418	1.837	-0.636	-1.265	-2	0.664
local	41	1.975	1.587	-0.293	-1.266	-2	0.9
local	37	2.206	1.755	0.373	-1.26	-2	0.873
local	118	2.165	1.697	1.137	-1.263	-2	0.839
local	50	1.395	1.105	-0.046	-1.265	-2	0.938
local	42	2.239	1.75	0.333	-1.266	-2	0.881
local	38	2.42	1.908	0.667	-1.261	-2	0.845
local	30	1.914	1.519	-0.046	-1.271	-2	0.908
local	57	1.978	1.547	1.108	-1.266	-2	0.85
local	39	1.896	1.455	1.193	-1.266	-2	0.828
local	34	2.143	1.687	1.535	-1.261	-2	0.397
local	15	1.806	1.428	1.114	-1.262	-2	0.787
local	27	2.063	1.627	0.396	-1.266	-2	0.895
local	26	2.43	1.937	1.251	-1.267	-2	0.855
local	17	2.26	1.78	0.921	-1.263	-2	0.875
local	11	1.879	1.504	0.339	-1.262	-2	0.864

Table 10: Results from daily version of the Plauborg model for 10cm depth.

scope	spesific scope	RMSE °C	MAE°C	bias °C	log(κ(model))	digit sensitivity	R^2
global	—	4.504	3.474	2.487	-0.796	-1	0.308
region	Østfold	4.348	3.363	1.901	-0.796	-1	0.424
region	Vestfold	4.564	3.47	2.297	-0.796	-1	0.397
region	Trøndelag	4.438	3.508	3.175	-0.796	-1	-0.194
region	Innlandet	4.688	3.568	2.601	-0.796	-1	0.353
local	52	3.556	2.841	0.559	-0.796	-1	0.604
local	41	4.248	3.286	1.677	-0.796	-1	0.491
local	37	4.754	3.675	2.174	-0.796	-1	0.391
local	118	4.726	3.654	3.208	-0.796	-1	0.162
local	50	4.048	3.025	2.207	-0.796	-1	0.434
local	42	4.863	3.741	2.364	-0.796	-1	0.393
local	38	4.832	3.682	2.601	-0.796	-1	0.308
local	30	4.465	3.433	2.015	-0.796	-1	0.456
local	57	4.655	3.636	3.153	-0.796	-1	0.125
local	39	4.31	3.39	3.083	-0.796	-1	0.081
local	34	4.583	3.675	3.471	-0.796	-1	-2.248
local	15	4.198	3.342	3.006	-0.796	-1	-0.241
local	27	4.672	3.547	2.535	-0.796	-1	0.415
local	26	5.17	4.009	3.282	-0.796	-1	0.302
local	17	5.049	3.84	2.939	-0.796	-1	0.336
local	11	3.821	2.924	1.692	-0.796	-1	0.35

Table 11: Results from the linear regression model for 20cm depth.

scope	spesific scope	RMSE °C	MAE°C	bias °C	log(κ(model))	digit sensitivity	R^2
global	—	4.231	3.267	2.303	-0.638	-1	0.423
region	Østfold	4.236	3.28	2.015	-0.638	-1	0.45
region	Vestfold	4.277	3.26	2.019	-0.638	-1	0.517
region	Trøndelag	4.133	3.274	2.893	-0.638	-1	0.028
region	Innlandet	4.282	3.254	2.246	-0.638	-1	0.504
local	52	3.679	2.889	1.226	-0.638	-1	0.221
local	41	3.976	3.07	1.494	-0.638	-1	0.593
local	37	4.501	3.503	2.07	-0.638	-1	0.473
local	118	4.5	3.486	2.93	-0.638	-1	0.306
local	50	3.611	2.702	1.766	-0.638	-1	0.584
local	42	4.571	3.525	2.109	-0.638	-1	0.506
local	38	4.815	3.741	2.502	-0.638	-1	0.388
local	30	4.053	3.106	1.733	-0.638	-1	0.588
local	57	4.293	3.356	2.775	-0.638	-1	0.295
local	39	4.057	3.193	2.835	-0.638	-1	0.213
local	34	4.262	3.415	3.176	-0.638	-1	-1.386
local	15	3.918	3.141	2.799	-0.638	-1	-0.003
local	27	4.272	3.236	2.078	-0.638	-1	0.551
local	26	4.714	3.651	2.902	-0.638	-1	0.456
local	17	4.518	3.432	2.506	-0.638	-1	0.501
local	11	3.567	2.713	1.529	-0.638	-1	0.51

Table 12: Results from the linear regression model for 10cm depth.

Norges miljø- og biovitenskapelige universitet
Noregs miljø- og biovitskapelege universitet
Norwegian University of Life Sciences

Postboks 5003
NO-1432 Ås
Norway