DEVELOPING A MACHINE LEARNING FRAMEWORK FOR ESTIMATING SOIL MOISTURE WITH VNIR HYPERSPECTRAL DATA

S. Keller\textsuperscript{1}, F. M. Riese\textsuperscript{1}, J. Stötzer\textsuperscript{1}, P. M. Maier\textsuperscript{1}, S. Hinz\textsuperscript{1}

\textsuperscript{1}Institute of Photogrammetry and Remote Sensing, Karlsruhe Institute of Technology, Karlsruhe, Germany
(sina.keller, felix.riese, johanna.stoetzer, philipp.maier, stefan.hinz)@kit.edu

Commission I, WG I/1

KEY WORDS: Hyperspectral data, Machine learning, Regression, Soil moisture, VNIR, Field campaign

ABSTRACT:

In this paper, we investigate the potential of estimating the soil-moisture content based on VNIR hyperspectral data combined with LWIR data. Measurements from a multi-sensor field campaign represent the benchmark dataset which contains measured hyperspectral, LWIR, and soil-moisture data conducted on grassland site. We introduce a regression framework with three steps consisting of feature selection, preprocessing, and well-chosen regression models. The latter are mainly supervised machine learning models. An exception are the self-organizing maps which combine unsupervised and supervised learning. We analyze the impact of the distinct preprocessing methods on the regression results. Of all regression models, the extremely randomized trees model without preprocessing provides the best estimation performance. Our results reveal the potential of the respective regression framework combined with the VNIR hyperspectral data to estimate soil moisture measured under real-world conditions. In conclusion, the results of this paper provide a basis for further improvements in different research directions.

1. INTRODUCTION

Precise data about spatial distributions and dynamics of soil moisture is valuable in many scopes of environmental applications. Hydrological as well as meteorological processes are influenced by soil moisture. Besides soils, microbes, and plants depend heavily on it (Farrelly et al., 2011; Cavagnaro, 2016; Tian et al., 2018). Apart from this, soil moisture emerges as one of the key variables relating to hydrological disasters such as flash floods on a catchment scale (Gill et al., 2006). Soil-moisture data of e.g. catchments areas functions as input variable for estimating and mitigating flood impacts to enhance flood models (Massari et al., 2014). In many regions, the soil-moisture distribution varies during the season of a year. Thus, e.g. summer soil-moisture availability serves as a relative indicator of a potential rate of fire spread, fire intensity, and fuel consumption (Girardin and Wotton, 2009). In addition, measured soil-moisture data is used to model the germination of seedbeds after such wildfires (Flrcheringer and Hardegree, 2004). Other studies have been conducted which refer to the linkages between soil moisture and wind erosion (Wang et al., 2014). All these fields of application have in common that they require soil-moisture estimations under almost real-world conditions such as a soil surface covered with vegetation.

The demand for spatial coverage and temporal resolution of soil moisture varies widely in the fields of application. Small-scale measurements, e.g. field site scale (pedon-scale), are performed with handheld sensors in combination with point-wise in situ soil-moisture measurements. One advantage of this scale is a high temporal resolution. Large-scale observations rely on airborne and satellite-based remote sensing solutions (cf. Maggioni et al., 2006; John, 1992; Finn et al., 2011; Colini et al., 2014). Therefore, they cover catchments and larger areas with a limited temporal resolution. Hence, a gap between the spatial coverage and temporal soil-moisture resolution as well as spatial coverages occurs (Robinson et al., 2008).

Developments in hyperspectral remote sensing during the last four decades have enhanced the data acquisition regarding e.g. spectral resolution for evaluating the soil-moisture dynamics. Terrestrial hyperspectral remote sensing sensors mounted on drones can cover a pedon-scale and are able to retrieve spectral signatures of the soil-moisture distribution in-between the top-soil layers (Kaleita et al., 2005). The surface of such sites is characterized by inhomogeneous covers including different vegetation, soil, and rock. This inhomogeneity of the soil surface results in overlaying reflectance spectra and poses a challenge to identify the soil-moisture state and dynamic (Salisbury and D’Aria, 1992). Since the datasets dealing with soil-moisture content are conducted expensively in field campaigns or laboratory measurements, most of them are of limited size.

When it comes to the estimation or modeling of soil-moisture contents based on remote sensing data, two trends can be deduced generally. First, hyperspectral sensors, which combine a fine spatial resolution and narrow bandwidths, outperform multispectral-retrieved data, especially in heterogeneous areas. Second, the short-wave infrared (SWIR) hyperspectral sensors obtain better results in estimating soil-moisture contents than the visible and near infrared (VNIR) sensors (Dalal and Henry, 1986; Finn et al., 2011). Crucial disadvantages of the SWIR sensors are the high acquisition cost, the need for active cooling, and, as direct consequence, the large weight and complex handling when mounting on e.g. a drone. Referring to these barriers and despite the knowledge of the great potential of such SWIR sensors, we seek to address the estimation of soil-moisture dynamics based on hyperspectral data in the wavelength of 450 nm to 950 nm (VNIR\textsuperscript{1} ). Furthermore, long-wavelength infrared (LWIR) data measured with an thermal camera is used.

For our investigations, we chose a dataset which has been mea-

\textsuperscript{1}We refer to this range of wavelength as visible and near infrared (VNIR) range due to reasons of simplicity.
soil-moisture estimation based on hyperspectral data.

and give an overview about future applications of the pedon-scale

of the methods follows in Section 4. In Section 5, we evaluate

machine learning models of the framework. The presentation

machine learning methods in Section 2. Subsequently, we de-

ing soil moisture based on hyperspectral data with and without

domain-knowledge like spectral information of vegetation.

The main contributions of this paper are:

• a detailed investigation of the potential of VNIR hyperspec-

tral data combined with LWIR data to estimate subsurface

soil moisture;

• an appropriate regression framework based on ten regres-

sion models such as partial least square (PLS), an arti-

ficial neural network (ANN), and a self-organizing map

(SOM) framework which merges unsupervised and super-

vised learning;

• a comprehensive evaluation of the regression performance

and an analysis of the potential of the underlying sensor data

for the estimation of subsurface soil-moisture dynamics on

a field site scale in regards to hydrological application.

We give a short overview on related work in regards to estimat-

ing soil moisture based on hyperspectral data with and without

machine learning methods in Section 2. Subsequently, we de-

scribe the measured dataset used for the evaluation of the several

machine learning models of the framework. The presentation

of the methods follows in Section 4. In Section 5, we evaluate

the proposed machine learning models. Finally, we conclude our

studies in Section 6, respond to the overlying regression problem

and give an overview about future applications of the pedon-scale

soil-moisture estimation based on hyperspectral data.

2. RELATED WORK

Traditionally, soil-moisture as well as pedo-hydrological dynam-

ics and states are monitored with point-based in situ measure-

ments using e.g. time domain reflectometry (TDR) probes and
tensiometers. Temporally high-resolution data can be aggregate
based on these sensors. The advantages of these techniques are
the precise measurement of the vertical soil-moisture distribution
at specific point locations. However, to obtain area-wide insight,
the traditional efforts are limited, time-consuming, and, depend-
ing on the experimental setup, uncertain (Jackisch et al., 2017).

At this point, the employment of hyperspectral remote sensing

techniques, covering the visible and near-infrared (VNIR), near-
infrared (NIR), short-wave infrared (SWIR), and the LWIR range

comes into effect. The performance to estimate soil moisture

based on VNIR, NIR and SWIR data enhances with increasing
wavelengths (Finn et al., 2011). The data acquisition with hyper-
spectral sensors ranges from point measurements with spectrora-
diometers to snapshots recordings by (drone-compatible) sensors

or satellites. The former provides a high spectral resolution, the
latter advantages area-wide recordings. Referring to Haubrock
(2008), only few studies investigating surface soil moisture via
airborne or spaceborne platforms record optical reflectance data.

Two distinct approaches are explored in regards to the estima-
tion of soil-moisture contents especially with hyperspectral data.
The first approach focusses on engineering features by combing
specific spectral bands to perform a ratio-calculation (Vereecken
et al., 2014; Fabre et al., 2015; Oltra-Carri et al., 2015).
The second approach relies on data-driven machine learning models
which develop their potential when handling non-linear regres-
sion problems or processing large datasets like in case of satellite-
based hyperspectral data (Guanter et al., 2015). Most machine
learning models are based on supervised learning such as par-
tial least square (PLS) regression, random forest (RF), support
vector machine (SVM), or artificial neural networks (ANN).

In addition, Riese and Keller (2018) introduce a framework of self-
organizing maps for the regression of soil moisture which com-
vides unsupervised and supervised learning.

According to the results of the feature engineering approaches
(first approach), the SWIR spectrum includes the most impor-
tant wavelengths which respond to soil-moisture contents (Dalal
and Henry, 1986; Wang et al., 2007; Haubrock, 2008; Finn et al.,
2011). A detailed review of modeling biomass and soil mois-
ture with several remote sensing data and inter alia with ma-
chine learning is stated in Ifarraguerri and Chang (2000). Fur-
ther remote sensing data such as C-band polarimetric SAR or
microwave radiometry is also applied to estimate soil moisture
in combination with machine learning (Baghdadi et al.,
2012; Pasolli et al., 2014; Xie et al., 2014). These datasets are
primarily conducted from satellite or airborne missions.

Generally, hyperspectral sensors provide spectral knowledge of
surface conditions. The soil surface represents a key factor for
the partitioning and redistributing of any precipitation before in-
filtrating into the subsurface (Jarvis, 2007; Brooks et al., 2015).
Subsurface soil-moisture dynamics and states are estimated based
on this spectral surface data combined with appropriate machine
learning models. Obviously, the spectral surface data repre-
sents an indirect approximation of the underlying physical soil-
moisture processes. Therefore, arising approximation uncertain-
ties add to the yet existing model uncertainties. In sum, the ben-

fits of hyperspectral applications prevail.

3. SENSORS AND DATASET

To evaluate the potential of VNIR hyperspectral sensors as in-
put data for the estimation of soil moisture, we rely on a dataset
which was conducted during a multi-sensor field campaign in
August 2017 in Linkenheim-Hochstetten, Germany. In this
pedon-scale field campaign, the vegetated surface as well as soil-
moisture states and dynamics have been monitored precisely.
Since real-world conditions are sustained, the ability to transfer
the applied regression methods is ensured. A detailed overview
of the field campaign with respect to the measurement setup as
well as its constraints and the analysis of the pedo-hydrological
processes can be found in Keller et al. (2018). Eight plots of an
undisturbed grassland site on loamy sand are the centerpiece of
the campaign. Figure 1a shows the plot setup. Each of these
plots covers an area of one square meter and is irrigated accord-
ing to a defined schema of various pulses (cf. Keller et al., 2018).
Multiple time domain reflectometry (TDR) probes measure the
value. The complete dataset consists of 1332 high-dimensional image data as input vector and the soil-moisture data as target value. The regression is performed with the hyperspectral and LWIR regression framework.

4.1 Feature selection

To estimate soil moisture in various depths from 2.5 cm to 20 cm. Based on these sensors, pedo-hydrological states and dynamics after the irrigation processes are surveyed. We refer to the TDR sensors in 5 cm depth as soil-moisture reference and ground truth within the scope of the paper.

A Cubert\textsuperscript{2} UHD 285 hyperspectral snapshot sensor records the hyperspectral image data (cf. Figure 1b). The measured reflectance includes the spectral signatures of among others the soil surface covered with vegetation. The hyperspectral sensor is installed on a stage at 10 m distance to cover the entire test site with one snapshot. Each hyperspectral snapshot contains 50 × 50 pixels and 125 spectral channels ranging from 450 nm to 950 nm with a spectral resolution of 4 nm. The pan-sharpened 1000 × 1000 pixels image in Figure 1b only serves as improved visualization of the measurement area, we use the raw hyperspectral image for the regression framework. As shown in Figure 1b, the measurement angles differ between the eight plots in the field of view of the hyperspectral sensor due to the necessary setup of the whole field campaign (cf. Keller et al., 2018). All reflectance spectra in every image are normalized based on a white reference resulting in reflectance values between 0 and 1. The spectralon as white reference is positioned visually in each snapshot to ensure this normalization after the recording. A thermal camera without active cooling (FLIR\textsuperscript{3} Tau 2 640) records the LWIR images (cf. Figure 1c) and is installed next to the hyperspectral camera. The LWIR images consist of 640 × 512 pixels, each characterized by a temperature value in °C. With respect to the approximated position of the TDR probes in the subsurface, average spectra of each plot and recording are calculated for both remote sensing data.

4.2 Preprocessing

Aiming to estimate soil-moisture values based on hyperspectral and one-dimensional LWIR input data, we foster the regression by applying two distinct preprocessing methods. The first method is a Principle Component Analysis (PCA) to reduce the dimensionality of the hyperspectral and LWIR input data (cf. Figure 2). It is applied to a stack of the VNIR reflectance values and the LWIR value. We use the first 20 principal components for the regression, since they cover most of the dataset variances. As second method, we apply a min-max scaling (cf. Figure 2). The scaling normalizes the input data to a fixed range between 0 and 1. In contrast to the PCA, the min-max scaling uses all input data, including the soil-moisture values.

During the preprocessing step, we pick either the PCA for dimensionality reduction, the min-max scaling for normalization purposes, or no preprocessing (cf. Figure 2, 2\textsuperscript{nd} step, left) is performed. We refer to the regression without preprocessing as the baseline framework. Later, in the test phase of the regression model, the results of the estimation based on each preprocessing method is compared against the baseline prediction result without any preprocessing.

4.3 Regression models

To estimate soil moisture, we select appropriate regression models and include them to the framework (cf. Figure 2, 3\textsuperscript{rd} step). These are linear regression (least-squares), partial least squares (PLS), random forest (RF), extremely randomized trees (ET), adaptive boosting (AdaBoost), gradient boosting (GB), k-nearest-neighbors (k-NN), support vector machines (SVM), artificial neural networks (ANN), and a framework of self-organizing maps.

Figure 1. An example of (a) an RGB image, (b) a hyperspectral snapshot, (c) an LWIR image in false colors. The hyperspectral snapshot of 50 × 50 pixels is pan-sharped to 1000 × 1000 only to improve visualization.

\textsuperscript{2}Cubert GmbH, Ulm, Germany
\textsuperscript{3}FLIR Systems. Inc., Portland, USA
1. Feature selection

![Feature selection diagram]

2. Preprocessing

- None
- OR
- PCA
- OR
- Min-Max scaler

3. Regression model

- Linear
- PLS
- RF
- ET
- AdaBoost
- GB
- SOM
- k-NN
- SVM
- ANN

During the training phase, the regression models are trained on the training subset by linking the hyperspectral and LWIR data to the soil-moisture target values. Except for the SOM, all regressors perform the training phase exclusively supervised. The SOM model includes two self-organizing maps to solve the regression problem, combining an unsupervised SOM with a supervised SOM. Riese and Keller (2018) introduce the schema of this SOM model.

The parameters of a regression model are divided into hyperparameters and model parameters. Model parameters are adapted during the training phase while hyperparameters are chosen beforehand. The optimal setup of the hyperparameters changes depending on the preprocessing methods in step 2 of the regression framework. Table 2 shows exemplarily the setup of the hyperparameters for the baseline framework (no preprocessing in step 2). We obtain a basic grid search with 10-fold cross validation on the training subset for each preprocessing method and regression model.

During the subsequent test phase, the trained regression framework estimates soil moisture on the basis of the hyperspectral and LWIR data of the test subset. The estimated soil-moisture values (model predictions) are compared to the measured soil-moisture values. The coefficient of determination \( R^2 \) and the root mean squared error (RMSE) express the regression performance. Since the framework in general relies on randomization, we obtain all regression results by seven independent training procedures each with different random seeds. The ensemble models RF, ET, AdaBoost, and GB provide additional information regarding the importance of the input variables (feature importance).

5. RESULTS AND DISCUSSION

By applying PCA-based dimensionality reduction, we obtain regression results relying on the first 20 principle components. By applying the min-max scaling to normalize the input data, the regression models rely on features in the range of 0 to 1. Using ten regression models with supervised or the combination of unsupervised and supervised learning principles for estimating soil moisture, the respective results are depicted in Table 1.

Both linear regression models (linear and PLS regressors) perform the worst. They are incapable of adapting to the high-dimensional regression problem. Within the ensemble models, RF and ET achieve good regression results. ET as an extension of the RF provides the best performance without preprocessing. An example of the relationship between the estimated and measured soil-moisture values of the ET regressor is given in Figure 4. GB estimates soil moisture slightly better than the AdaBoost. The influence of min-max scaling on ensemble models is negligible. Figure 5 shows the feature importance of the input variables of the baseline framework provided by the ensemble models. As expected, RF and ET as averaging ensemble models prioritize similar features (input variables).
The distributions of the feature importance of both boosting models AdaBoost and GB differ. Therefore, the linkages between the spectral features and the underlying physical processes ask for a detailed further analyses. Strong correlations between the input features (hyperspectral spectral and the LWIR values) appear more challenging for the performance of both boosting models than for the performance of the averaging models.

In addition, Figure 6 shows a combination of the hyperspectral mean spectrum with the standard deviation and the feature importance of the ET regressor. On the right side, the LWIR data and its importance for the regression is illustrated. The spectral bands in the area of 826 nm are significantly more important than the remaining ones. While the variance of the mean spectrum is relatively large in the upper third of the spectrum, the feature importance distribution continually decreases after the identified peak. Bands between 890 nm to 930 nm possess a minor feature importance and a large variance. This finding indicates that these bands exhibit noise, e.g. occurring due to weather and sensor conditions. The feature importance of the LWIR value plays a minor role.

The k-NN model generally works well and improves with min-max-scaled data due to the linkage between the integrated distance measure and the normalized data. With respect to the SVM, ANN, and the SOM models, the same effects apply.

Considering the complete performance of the regression framework, we state that the preprocessing with the PCA-based dimensionality is insufficient of solving the present regression problem. The normalization with the min-max scaling seems more favorable for estimating soil moisture based on hyperspectral and LWIR data. This scaling yields the best regression results for almost any regression model. We can address an additional preprocessing method by combining a min-max scaling and a PCA as well as other preprocessing techniques and their effects on the regression performance in further studies. We would like to point out that improving the tuning process of the hyperparameters could further enhance the regression results.

In addition, we remark that the regression framework is data-driven. It estimates soil-moisture values based on pure reflectance spectra without relying on additional information, e.g. vegetation spectra and information of the measurement angle. Another notable aspect of estimating soil moisture appears when focusing on the spatial soil-moisture distribution. This distribution highly depends on further factors such as coverages with mixed vegetation or soil structure. Thus, it could be extremely inhomogeneous even within small areas. Furthermore, we would like to take a glance look on the accuracy of the measured soil-moisture reference data. According to specification of the installed TDR sensors, their soil-moisture measuring accuracy varies between 1 p.p. to 2 p.p. depending on the moisture values. Such measurement errors result in a number of effects with respect to the regression framework and finally to the estimation performance. These effects should be investigated in further work.

6. CONCLUSION

In this paper, we address the estimation of soil moisture based on a measured, pedon-scale dataset. In contrast to most datasets applied in the context of estimating soil moisture with hyperspectral data, the underlying data consists of VNIR hyperspectral data combined with LWIR data. The hyperspectral data includes spectral signatures of a vegetated soil surface to ensure an application under real-world conditions prevailing e.g. at a catchment area. Our main objective is to investigate the potential of solving the regression problem solely with this data.

We introduce an appropriate regression framework involving one (optional) preprocessing step and nine supervised regression models as well as one model which combines an unsupervised SOM and a supervised SOM. The results of the regression framework reveal the potential of respective data-driven models in combination with the used input data under varying real-world measurement circumstances. In this context, machine learning provides a data-driven solution without exclusively relying on domain knowledge.

The following challenges are mastered satisfactorily:

- the limited size of data,
- their VNIR spectrum range which is suboptimal referring to preceding studies,
- the fact that we estimate soil moisture with an actively vegetated surface which also is suboptimal to preceding studies, and
- the measurement angles differing for each plot.

To conclude, we point out that it is possible to retrieve soil-moisture content from measured VNIR hyperspectral data due to the outweigh of the benefits.

As a direct consequence, we will approach further improvements in different research directions which we point out in the discussion section (cf. Section 5). In future work, we plan to analyze in detail the impacts of inhomogeneous soil-moisture distributions and the error propagation which starts with the soil-moisture measuring accuracy of the sensors and relying on this data as reference. Thereby, we also intend to conduct a dataset on an analog field experiment but using a SWIR sensor for the reflectance measurements. Then, we are able to evaluate the performance of the presented regression framework in this dataset and are able to compare the performance with SWIR and VNIR input data.

References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., Kudlur, M., Levenberg, J.,
### Table 1. Regression results for the soil-moisture estimation.

| Model   | $R^2$ baseline | RMSE in % soil moisture | $R^2$ with PCA | RMSE in % soil moisture | $R^2$ with scaling | RMSE in l |
|---------|----------------|-------------------------|----------------|-------------------------|--------------------|-----------|
| Linear  | 50.7           | 2.5                     | 49.3           | 2.6                     | 50.7               | 0.1       |
| PLS     | 52.0           | 2.5                     | 49.3           | 2.6                     | 48.3               | 0.1       |
| RF      | 67.0           | 2.1                     | 63.2           | 2.2                     | 66.9               | 0.1       |
| ET      | **73.0**       | **1.9**                 | **69.1**       | **2.0**                 | **72.8**           | **0.1**   |
| AdaBoost| 59.6           | 2.3                     | 55.2           | 2.4                     | 56.4               | 0.1       |
| GB      | 65.2           | 2.1                     | 58.8           | 2.3                     | 65.3               | 0.1       |
| k-NN    | 53.5           | 2.5                     | 53.8           | 2.5                     | 72.5               | 0.1       |
| SVM     | 50.7           | 2.5                     | 50.2           | 2.6                     | 70.4               | 0.1       |
| ANN     | 32.9           | 2.9                     | 52.3           | 2.5                     | 60.1               | 0.1       |
| SOM     | 42.5           | 2.7                     | 43.0           | 2.7                     | 56.5               | 0.1       |

Figure 6. Mean spectrum with the standard deviation as vertical error bars of the hyperspectral data (blue) and the LWIR data (red) of the complete dataset. The feature importance of the ET regressor is shown in orange.
Dalal, R. C. and Henry, R. J., 1986. Simultaneous determination of mois-
ture, organic carbon, and total nitrogen by near infrared reflectance
spectrophotometry. *Soil Science Society of America Journal* 50(1), pp.
120–123.

Finn, M. P., Lewis, M., Bosch, D. D., Giraldo, M., Yamamoto, K., Sul-
livon, D. G., Kincaid, R., Luna, R., Allam, G. K., Kvien, C. and Wil-
kins, M. S., 2011. Remote sensing of soil moisture using airborne hyperspectral data. *G/Science & Remote Sensing* 48(4), pp.
522–540.

Flerchinger, G. and Hardegree, S., 2004. Modelling near-surface soil tem-
perature and moisture for germination response predictions of post-
wildfire seedbeds. *Journal of Arid Environments* 59(2), pp.
369–385.

Freund, Y. and Schapire, R., 1997. A decision-theoretic generalization of
on-line learning and an application to boosting. *55(1)*, pp. 119–139.

Friedman, J., Hastie, T. and Tibshirani, R., 2001. The elements of statis-
tical learning. Vol. 1, Springer series in statistics New York.

Geurts, P., Ernst, D. and Wehenkel, L., 2006. Extremely randomized
trees. *Machine Learning* 63(1), pp. 3–42.

Gill, M. K., Asefa, T., Kemblowski, M. W. and McKee, M., 2006. Soil
moisture prediction using support vector machines. *JAWRA Journal of
the American Water Resources Association* 42(4), pp. 1033–1046.

Girardin, M. P. and Wotton, B. M., 2009. Summer moisture and wildfire
risks across canada. *Journal of Applied Meteorology and Climatology*
48(3), pp. 517–533.

Guanter, L., Kaufmann, H., Segl, K., Foerster, S., Rogass, C., Chabril-
lat, S., Kuester, T., Hollstein, A., Rossner, G., Chlebek, C., Straif, C.,
Fischer, S., Schrader, S., Storch, T., Heiden, U., Mueller, A., Bach-
mann, M., Mile, H., Miller, R., Habermeyer, M., Ohndorf, A., Hii, J.,
Buddenbaum, H., Hostert, P., Linden, S. v. d., Leito, P. J., Rabe, A.,
Doeffer, R., Krasemann, H., Xi, H., Mauser, W., Hank, T., Locherer,
M., Rast, M., Staenz, K. and Sang, B., 2015. The EnMAP Spaceborne
Imaging Spectroscopy Mission for Earth Observation. *Remote Sensing*
7(7), pp. 8830–8857.
APPENDIX

The appendix contains the setup of the hyperparameters for all regression models without preprocessing, cf. Table 2.

Table 2. Hyperparameter setup for the regression framework without preprocessing. This setup is obtained by a basic grid search algorithm with 10-fold cross validation on the training subset. The regressors are implemented mostly in scikit-learn (Pedregosa et al., 2011) and TensorFlow (Abadi et al., 2016), while the SOM is implemented according to Riese and Keller (2018).

| Model   | Reference                          | Package       | Hyperparameter setup                                                                 |
|---------|------------------------------------|---------------|--------------------------------------------------------------------------------------|
| Linear  | –                                   | scikit-learn  | –                                                                                   |
| PLS     | –                                   | scikit-learn  | \(n_{\text{components}} = 10; \text{max}_\text{iter} = 100; \text{tol} = 10^{-7}\) |
| RF      | Breiman (2001)                      | scikit-learn  | \(n_{\text{estimators}} = 1000\)                                                  |
| ET      | Geurts et al. (2006)                | scikit-learn  | \(n_{\text{estimators}} = 1000\)                                                  |
| AdaBoost| Freund and Schapire (1997)          | scikit-learn  | learning rate = 3.0; \text{loss} = "linear"; \(n_{\text{estimators}} = 150\)       |
| GB      | Breiman (1997)                      | scikit-learn  | learning rate = 0.1; \text{loss} = "huber"; \(n_{\text{estimators}} = 1000; \text{max}_\text{depth} = 2\) |
| k-NN    | Altman (1992)                       | scikit-learn  | \(n_{\text{neighbors}} = 6; \text{weights} = "distance"; \text{leaf}_\text{size} = 1\) |
| SVM     | Vapnik (1995)                       | scikit-learn  | \(C = 26827; \gamma = 0.00178\)                                                  |
| ANN     | Friedman et al. (2001)              | TensorFlow    | Keras sequential model with epochs = 70; batch size = 8; four dense layers with \{64, 128, 64, 32\} neurons and RELU activations |
| SOM     | Kohonen (1990); Riese and Keller (2018) | other        | SOM size = 30 \times 70; \(N_{\text{Iterations, Input}} = 5000; N_{\text{Iterations, Output}} = 8000\); learning rates \(\alpha_{\text{Start}} = 0.4; \alpha_{\text{End}} = 0.005\); exponential neighborhood function (input and output); pseudo-gaussian neighborhood distance weight |