https://doi.org/10.5194/egusphere-2023-1872
Preprint. Discussion started: 31 August 2023

Potential of Machine learning techniques compared to MIKE-SHE model for drain flow predictions in tile-drained agricultural areas of Denmark

Hafsa Mahmood[1], Ty P. A. Ferré[2], Raphael J. M. Schneider[3], Simon Stisen[3], Rasmus R. Frederiksen[4], Anders V. Christiansen[1]

1. hm@geo.au.dk, anders.vest@geo.au.dk : Aarhus University, Department of Geoscience, Aarhus, Denmark

2. tyferre@arizona.edu : University of Arizona, Department of hydrology and atmospheric sciences, Tucson, Arizona, USA

3. rs@geus.dk, sst@geus.dk : Geological Survey of Denmark and Greenland (GEUS), Department of Hydrology, Copenhagen, Denmark

4. rrf@ecos.au.dk: Aarhus University, Department of Ecoscience, Aarhus, Denmark

Correspondence to: Hafsa Mahmood (hm@geo.au.dk)

Abstract

Temporal drain flow dynamics and understanding of their underlying controlling factors are important for water resource management in tile-drained agricultural areas. The use of physics-based water flow models to understand tile drained systems is common. These models are complex, with large parameter sets and require high computational effort. The primary goal of this study was to examine whether simpler, more efficient machine learning (ML) models can provide acceptable solutions.

The specific aim of our study was to assess the potential of ML tools for predicting drain flow time series in multiple catchments subject to a range of climatic and landscape conditions. The investigation is based on unique data containing time series of daily drain flow in multiple field scale drain sites in Denmark. The data include: climate (precipitation, potential evapotranspiration, temperature); geological properties (clay fraction, first sand layer thickness, first clay layer thickness); and topographical indexes (curvature, Topographical wetness indexes, Topographical position index, elevation). Both static and dynamic variables are used in the prediction of drain flows. The ML algorithm extreme gradient boosting (XGBoost) and convolutional neural network (CNN) were examined, and the results were compared with a physics-based distributed model (MIKE-SHE).

The results show that XGBoost performs similarly to the physics-based MIKE-SHE models, and both outperform CNN. Both ML models required significantly less effort to build, train, and run than MIKE-SHE. In addition, the ML models support efficient feature importance analysis. This showed that climatic variables were important for CNN models and XGBoost. The results support the use of ML models for hydrologic applications with sufficient data for training. Further, the insights offered by the feature

https://doi.org/10.5194/egusphere-2023-1872
Preprint. Discussion started: 31 August 2023

importance analysis may support further data collection and developments of physics-based models when existing data are insufficient to support ML approaches.

1. Introduction

Tile drain flow prediction is important for sustainable water resource management because tile drains are crucial for accurate quantification of subsurface water fluxes in tile drained fields, which has direct impacts on predicting adjacent surface water flow. Approximately half of the agricultural land in Denmark has subsurface drains (Moller et al., 2018). However, only a small fraction of these sites has tile drain flow monitoring. Because tile-drained fields can have very rapid communication with surface water and surface water pollution, it will become increasingly important to understand their hydrologic impacts under climate change (Golmohammadi et al., 2021; Jeantet et al., 2022).

Physics-based models, both distributed and lumped/conceptual, have been used for predicting drain flows. These include: HSPF (Singh et al., 2005), MIKE SHE (De Schepper et al., 2017; Hoang et al., 2014; Mahmood et al., 2023), CATHY (Muma et al., 2014), SWAT (Hoang et al., 2014; Singh et al., 2005), RZWQM (Craft et al., 2018), DRAINMOD (Northcott et al., 2001; Youssef et al., 2021), Hydro Geosphere (De Schepper et al., 2017) and MODFLOW (Mirlas, 2009). In some cases, physics-based distributed models can predict drain flows accurately, but they are time and data-intensive to build and calibrate (Basha et al., 2008; Beven, 1989). In contrast, lumped models require relatively less data and are more computationally efficient. But the unclear physical meaning of the parameters reduces these to correlative models with limited transferability. The question addressed in this study was whether ML models, which are also generally considered to be correlative, are more efficient than either distributed or lumped physics-based models (Herath et al., 2021).

There is a growing interest in the application of ML models in hydrology (Shen, 2018). Most of these studies are aimed at the prediction of the water table depth (Koch et al., 2019; Koch et al., 2021; Schneider et al., 2022), water quality (Erickson et al., 2021; Tesoriero et al., 2015) and stream flow (Bechtold et al., 2014; Kratzert et al., 2019; Kuzmanovski et al., 2015; Mushtaq et al., 2022; Xu et al., 2020; Zia et al., 2015). Few studies addressed the prediction of drain flows (Bjerre et al., 2022; Frederiksen et al., 2023; Kuzmanovski et al., 2015; Motarjemi et al., 2021). Kuzmanovski et al. (2015) predicted surface runoff and tile drain flow on a single agricultural field. Motarjemi et al. (2021) predicted only the cumulative annual tile drain flows over multiple catchments. Bjerre et al. (2022) focused on spatial drain flow predictions rather than timeseries. The current study considers high temporal resolution drain flow prediction over long time series for multiple sites. The results are compared with a physics-based model and insights are extracted from feature importance analysis.

Two different ML methods are commonly used: decision trees (Bechtold et al., 2014; Bjerre et al., 2022; Erickson et al., 2021; Koch et al., 2019; Koch et al., 2021; Kuzmanovski et al., 2015; Mushtaq et al., 2022; Schneider et al., 2022; Zia et al., 2015) and neural networks (Dai et al., 2023; Koch & Schneider, 2022; Lees et al., 2022; Motarjemi et al., 2021; Xu et al., 2020). The performance of these approaches are compared, which in the context of drain flow only has been attempted once previously: Motarjemi et al. (2021)

https://doi.org/10.5194/egusphere-2023-1872
Preprint. Discussion started: 31 August 2023

conducted a comparative study on the use of multiple machine learning methods for annual tile drain. Specifically, we tested several decision tree approaches and selected XGBoost, a gradient boosting technique, which has shown optimal performance in international competitions (Chen & Guestrin, 2016). Among the many available neural network methods, we opted for Convolutional Neural Networks (CNN) (Yao et al., 2019). We chose CNNs based on their applicability to the problem of tile-drained fields and based on previous research (Bai et al., 2018) that indicates that a 1D-CNN performed better in predicting timeseries than other sequence modelling neural networks such as long short-term memory (LSTM) and Gated recurrent unit (GRU).

The objective of this study was to investigate the potential of ML models for tile-drain flow prediction and evaluate their transferability to ungauged basins. The sub-objectives of this study were.

(1) Compare XGBoost and CNN for predicting daily drain flows for different catchments of Denmark.

(2) Compare the results of both ML techniques with an existing, physics-based model (MIKE-SHE) that has been calibrated on the same catchments.

(3) Identify the important observations (features) that contribute to the prediction of daily drain flow

To the best of our knowledge, this study is the first application of XGBoost and CNN models for predicting daily drain flows for multiple sites. To evaluate the transferability of the ML models, we employed the "leave one cluster out technique," where we tested and verified the models for each cluster individually. This involved training the models using data from all clusters except the one being tested. We conducted tests on four clusters with a total of 20 drain sites. We also had an additional four drain sites which were always part of the training dataset as they did not belong to any of the four clusters.

2. Methodology

2.1. Study site and target feature

In this study, a total of 24 field-scale drain sites were chosen from various locations in Denmark. Out of 24 drain sites, 20 belonged to four main clusters. All drain sites ranged from 1 to 100 ha in area (Table 1). For Gyldenholm, Lolland, and Lillebæk cluster data were available for four drain sites each. For Norsminde, cluster data were available for a total of eight sites. Another four sites included in the study were Vadum, Gedved, Fillerup and Ulvsborg (Figure 1). The focus of the study was to predict the daily drain flow (this is defined as the ‘target feature’). The drain flow timeseries data for each drain site is provided in the Table 1. The average daily drain flow across all drain sites was 0.51 mm/d with standard deviation of 1.17 mm/d and skewness of 5.75 mm/d. The positive skewness suggests that most of the data points are concentrated towards the lower end of the range, with a few extremely high drain flow values influencing the mean (Figure 2).

https://doi.org/10.5194/egusphere-2023-1872
Preprint. Discussion started: 31 August 2023

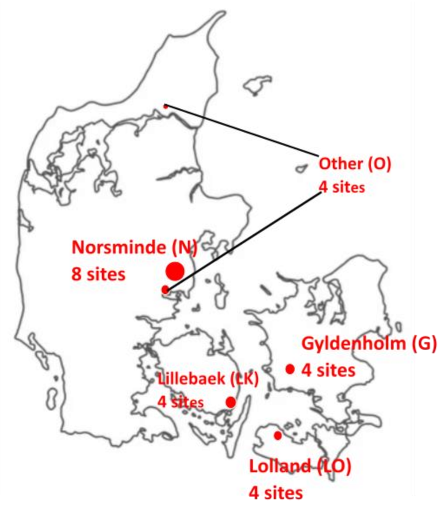

Figure 1 Location of drain catchments in Denmark

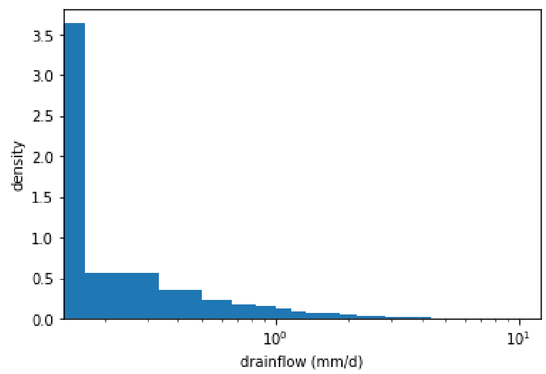

Figure 2 Distribution of drain flow data

https://doi.org/10.5194/egusphere-2023-1872
Preprint. Discussion started: 31 August 2023

Table 1 Description of clusters and their drain sites

Clusters	Drain site Id	Drain sites	Area (ha)	Drain time series start	Drain time series end	Model 1	Model 2	Model 3	Model 4
"other"	**O_1**	**Fillerup**	38.3	12-12-2012	29-05-2017	Train-test (5 kfold cross-validate)	Train-test (5 kfold cross-validate)	Train-test (5 kfold cross-validate)	Train-test (5 kfold cross-validate)
	O_2	**Ulvsborg**	34.9	30-11-2015	01-05-2018				
	O_3	**Gedved**	35.3	09-01-2020	15-05-2021				
	O_4	**Vadum**	8.9	02-07-2013	19-01-2017				
Norsminde	**N_1**	**Norsminde1**	34.0	20-04-2012	08-06-2017			Verify	
	N_2	**Norsminde2**	32.9	21-04-2012	18-06-2017				
	N_3	**Norsminde3**	27.5	22-04-2012	27-06-2017				
	N_4	**Norsminde4**	4.0	27-09-2012	27-06-2017				
	N_5	**Norsminde5**	11.9	27-09-2012	08-06-2017				
	N_6	**Norsminde6**	7.2	18-09-2012	06-02-2018				
	N_7	**Norsminde7**	3.6	18-09-2012	03-07-2019				
	N_8	**Norsminde8**	6.2	26-09-2012	13-12-2018				
Lillebaek	**LK_1**	**Lillebaek1**	1.0	01-01-1989	31-12-2020	Verify		Train-test (5 kfold cross-validate)	
	LK_2	**Lillebaek2**	4.5	01-01-1990	31-12-1999				
	LK_3	**Lillebaek3**	1.0	01-01-1990	31-12-1999				
	LK_4	**Lillebaek4**	2.6	01-01-1989	31-12-2018				
Lolland	**LO_1**	**Lolland1**	2.6	01-01-1989	01-01-2020	Train-test (5 kfold cross-validate)	Verify		
	LO_2	**Lolland2**	5.8	01-01-1989	01-01-2020				
	LO_3	**Lolland3**	2.6	01-01-1989	01-01-2019				
	LO_4	**Lolland4**	2.0	01-01-1994	01-01-2020				
Gyldenholm	**G_1**	**Gyldenholm1**	4.6	11-11-2015	09-05-2018		Train-test (5 kfold cross-validate)		Verify
	G_2	**Gyldenholm2**	48.6	11-11-2015	09-05-2018				
	G_3	**Gyldenholm3**	120.0	11-11-2015	09-05-2018				
	G_4	**Gyldenholm4**	33.6	11-11-2015	09-05-2018				

2.2. Feature selection and downscaling

A set of static and dynamic features were selected, referred to as the 'predictor variables'. These features included drain catchment properties, climate, topographical and geological features. They are listed in Table 2.

https://doi.org/10.5194/egusphere-2023-1872
Preprint. Discussion started: 31 August 2023

Table 2 Features used for prediction of drain flow.

Feature type	Feature name	Abbreviation	Mean (max, min)	Selected features
Dynamic features	***Climate variables***			
	Mean precipitation of the day (mm/d)	prec	2.05 (0, 33.19)	X
	Mean precipitation of previous day (mm/d)	prec_prev_day	2.05 (0, 33.19)	X
	Mean precipitation of previous week (mm/d)	Prec_prev_wk	2.05 (0, 17.50)	X
	Mean precipitation of previous month (mm/d)	prec_prev_mon	2.05 (0, 7.94)	X
	Mean temperature of the day (◦C)	temp	9.04 (-9.89, 25.22)	X
	Mean evapotranspiration of the day (mm/d)	eva	1.19 (0, 5.59)	X
	Mean precipitation of previous 6 months (mm/d)	prec_prev_6mon	2.05 (0.86, 3.97)	
	Mean precipitation of previous year (mm/d)	prec_prev_yr	2.05 (0.86, 3.97)	
	Mean evapotranspiration of previous day (mm/d)	eva_prev_day	1.19 (0, 5.59)	
	Mean evapotranspiration of previous week (mm/d)	eva_prev_wk	1.19 (0, 5.11)	
	Mean evapotranspiration of previous month (mm/d)	eva_prev_mon	1.19 (0.09, 3.67)	
	Mean evapotranspiration of previous 6 months (mm/d)	eva_prev_6mon	1.19 (0.37, 2.34)	
	Mean evapotranspiration of previous year (mm/d)	eva_prev_yr	1.19 (0.87, 1.52)	
Static features	***Topography variables***			
	Standard deviation of elevation (m)	elev_std	3.52 (0.17, 10.55)	X
	Mean of absolute TPI in 20m radius	tpi_20	0.15, (0.04, 0.34)	X
	Mean of absolute TPI in 200m radius	tpi_200	1.19 (0.13, 2.82)	X
	Mean of TWI	twi	9.96 (8.99, 11.12)	X
	Standard deviation of TWI	twi_std	1.13 (0.66, 1.39)	X
	Mean of absolute TPI in 20m radius around drain catchment in 300m buffer	around_cat_tpi_20	1.13 (0.16, 2.21)	X
	Mean of TWI around drain catchment in 300m buffer	around_cat_twi	9.84 (9.25, 10.75)	X
	Standard deviation of TWI in 20m radius around drain catchment in 300m buffer	around_cat_twi_std	1.27 (1.12, 1.54)	X
	Mean of absolute TPI in 10m radius	tpi_10	0.09 (0.03, 0.19)	
	Mean of absolute TPI in 50m radius	tpi_50	0.33 (0.05, 0.82)	
	Mean of absolute TPI in 100m radius	tpi_100	0.61 (0.07, 1.60)	
	Mean of absolute curvature	curvature	0.28 (0.11, 0.64)	
	Slope (degrees)	slope_degree	2.14 (0.26, 5.36)	
	Standard deviation of elevation around drain catchment in 300m buffer	outside_cat_elev_std	6.05 (0.43, 13.05)	
	Geology variables			
	Mean Clay content a, b, c horizon (%)	clay_content	14.21 (3.57, 21.02)	X
	Standard deviation of clay content a, b, c horizon (%)	clay_content_std	0.84 (0.30, 2.25)	X
	Clay thickness (m)	clay_thickness	40.55 (0.22, 388.95)	X
	Variance in clay thickness (m)	clay_thickness_var	0.11 (0.07, 0.16)	
	Clay a horizon (%)	clay_content_a	12.13 (3.70, 17.47)	
	Clay b horizon (%)	clay_content_b	14.86 (3.67, 26.04)	
	Clay c horizon (%)	clay_content_c	15.60 (2.96, 22.47)	
	Clay d horizon (%)	clay_content_d	15.77 (4.42, 21.39)	
	Other			
Static features	Drain catchment	area	$2.0x10^5$ ($9.7x10^3$, $1.2x10^6$)	X
Dynamic features	Hydrological day of the year	day_of_year		X
	Month	month	-	
	Quarter of year	quarter	-	

Climate features were obtained from the Department of Hydrology, Geological Survey of Denmark and Greenland, but originated from the Danish Meteorological Institute (DMI) as 10km/20km gridded daily data (DMI, 2023; Scharling, 1999a, 1999b). Precipitation (mm), evapotranspiration (mm) and temperature (◦C) were obtained for all drain catchments. For evapotranspiration and precipitation, other features such as mean previous day, mean previous week, mean previous month, mean previous 6 months, and mean previous year were also calculated and included as predictor variables.

Topographical and geological features were static features. The digital elevation model at 10m resolution was used to derive topographical features for each drain site including: standard deviation of elevation (m.s.l); absolute mean Topographical position index (TPI, in radii of 10m, 20m, 50m, 100m and 200m); absolute mean curvature; mean slope; and topographical wetness index

https://doi.org/10.5194/egusphere-2023-1872
Preprint. Discussion started: 31 August 2023

(TWI). We also calculated the standard deviation and absolute mean of TWI, standard deviation of elevation (m.s.l), and absolute mean TPI in a 300m buffer around each drain catchment.

Geological features included the drain catchment average clay content (%) in horizon a (0-5 cm depth), horizon b (5-15 cm), horizon c (15-30 cm), and horizon d (30-60 cm) as developed by Adhikari et al. (2013). Standard deviation and clay content (%) were also estimated across horizons a, b and c. The drain catchment average of first clay layer thickness and sand layer thickness from the nationwide hydrogeological interpretation (EPA, 2020) was obtained. Variance in clay thickness across each drain catchment was estimated.

Another static feature was the area of the drain catchment (m^2). Additional dynamic features included the hydrological day of year, quarter of year, and month. In total 39 features were included in the initial analysis. This set was reduced to 19 by developing a covariance matrix and removing any feature that had a Pearson R value above 0.85 with any other feature.

2.3. Xgboost

XGBoost stands for "Extreme Gradient Boosting". It is an enhanced version of boosting ensemble techniques, specifically designed for decision trees in the classification and regression trees (CART) family. Even though it is based on the gradient boosting framework, it is more scalable and has faster speed due to its suitability for parallel computing techniques and readily available hyperparameter optimization tools. XGBOOST operates by iteratively improving predictions in an additive manner. During each iteration, weak classifiers are employed, and their errors are used to enhance subsequent classifiers. Misclassified samples are given greater emphasis in subsequent steps, compelling the classifier to focus on improving their performance. The final classification result benefits from the collective improvement of all the previously built trees (Chen & Guestrin, 2016). We used python library 'xgboost' to make XGBoost models (Chen & Guestrin, 2016).

2.4. CNN

The 2D-Convolutional neural networks were initially developed for image recognition (Lawrence et al., 1997). However, 1D-CNN has been used for time series analysis (Lewinson, 2020). CNN model structure consists primarily of three different types of layers: convolutional layers; pooling layers; and fully connected layers. Convolutional layers allow feature extraction using a filter to produce multiple feature maps from given features. Moreover, a kernel size in the convolutional layer allows to set the number of previous timesteps used for a specific time. To introduce non-linearity, an activation function is used; in our case, we used a rectified linear unit (relu) in the convolutional layers. Pooling layers reduce the size of the time series by preserving the most important features identified by the convolutional layers. These pooled layers are used to build fully connected layers to map the extracted features. Fully connected layers are associated with the loss function that estimates the error between observed and predicted values (Lawrence et al., 1997). We used the python library Tensorflow to make CNN models (Abadi et al., 2016).

https://doi.org/10.5194/egusphere-2023-1872
Preprint. Discussion started: 31 August 2023

2.5. Machine learning model setup

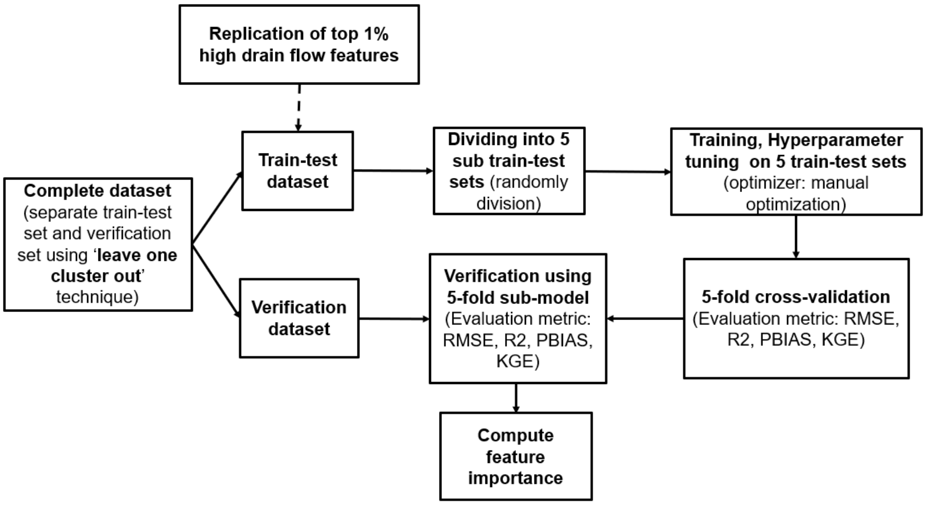

Figure 3 Workflow of both models

2.5.1. Division of train-test dataset and verification data

To test the potential of ML techniques, the data were split into two different subsets: the train-test dataset and the verification dataset. The train-test and verification datasets were defined using the 'leave one cluster out technique'. In this technique, we kept all the drain sites of one cluster to use as the verification dataset while training and cross-validation used the rest of the data. This technique was applied on each of the four clusters and leading to four CNN models and four XGBoost models.

2.5.2. High drain flow data replication and division into K-folds of train-test dataset

Once the train-test dataset was separated, we replicated multiple times the top 1% high drain flow values of the train-test dataset (see Table 3 for XGBoost models and Table 4 for CNN models). This replication was performed to increase the weight of the extremely high drain flow values as high drain flow values were rare in the dataset (Table A1). We aimed that the replication would improve the model performance. After replication of high drain flows data, the train-test dataset was divided into 5 subsets, each containing an equal number of samples. This splitting was done randomly. Four out of five subsets were used for training and the one left subset was used for testing. The process was repeated until all the subsets were tested separately.

2.5.3. Hyperparameter tuning and cross validation

In the 5-fold cross-validation, each model was trained and evaluated five times, each time using a different fold as the test set and the remaining four folds as the training set. The XGBoost and CNN models have different structures and, therefore, have different hyperparameters. Each of the 4 models also had different parameters. In each case, tuning involved jointly optimizing all 5 folds of each model simultaneously using python. Different built in loss functions were tested for model optimization including the root mean squared error (RMSE), mean absolute error (MAE), and root mean squared log error (RMSLE). The formulas of the following are given bellow:

https://doi.org/10.5194/egusphere-2023-1872
Preprint. Discussion started: 31 August 2023

Equation 1

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(Q_{sim\ i} - Q_{obs\ i})^2}{N}}$$

Equation 2

$$RMSLE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\log(\,Q_{sim\ i} + 1) - \log\,(Q_{obs\ i} + 1))^2}$$

Equation 3

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|Q_{sim\ i} - Q_{obs\ i}|$$

Where Q_{sim} is simulated drain flow and Q_{obs} is observed drain flow and N is the number of data points.

2.5.4. Evaluation metric:

Among the evaluation metrics, RMSE, Kling-Gupta cofficient (KGE), percentage bias (PBIAS) and coefficient of determination (R^2) were used. The formulas are given below:

Equation 4

$$R^2 = 1 - \frac{\sum_{i=0}^{N}(Q_{obs,i} - Q_{sim,i})^2}{\sum_{i=0}^{N}(Q_{obs,i} - Q_{obs\_mean})^2}$$

Here Q_{obs_mean} is mean of all observed drain flows.

Equation 5

$$KGE = 1 - \sqrt{(r-1)^2 + \left(\frac{\sigma_{sim}}{\sigma_{obs}} - 1\right)^2 + \left(\frac{\mu_{sim}}{\mu_{obs}} - 1\right)^2}$$

Here, r is the linear correlation between observed and simulated drain flows, σ_{sim} is the standard deviation of the simulated drain flows, σ_{obs} is standard deviation of the observed drain flow, μ_{sim} is mean of the simulated drain flows and $\mu_{obs}$is mean of the observed drain flows (Gupta et al., 2009).

Equation 6

$$PBIAS = \left[\frac{\sum_{i=1}^{N}\left(Q_{sim,i} - Q_{obs,i}\right) * 100}{\sum_{i=1}^{N} Q_{obs,i}}\right]$$

https://doi.org/10.5194/egusphere-2023-1872
Preprint. Discussion started: 31 August 2023

2.5.5. Model verification and feature importance

After optimizing the model hyperparameters, the drain flow predictions were performed individually for each of the 5-fold sub-models on the verification dataset (verification data shown in Table A1 for each model). Subsequently, the predicted drain flow results were combined or aggregated to obtain the consolidated or overall results. Moreover, we calculated the importance of each feature using the permutation method for all XGBoost and CNN models. The python library sklearn.inspection was used to find permutation importance. The permutation method randomly shuffles the values of a single feature and measures the resulting impact on the model's performance (Pedregosa et al., 2011). If shuffling the feature significantly decreases performance, it is considered important. By applying this technique to all the features individually, the relative importance of each feature can be quantified.

2.6. Existing National Hydrological Model of Denmark

The 10m resolution drain models were specially developed for drain flow prediction by Mahmood et al. (2023). It is a physics-based fully distributed model developed in MIKE-SHE. The model was jointly calibrated using drain flow data from the same drain sites used in this study except Gedved (O_3) because data for Gedved was obtained at a later stage of the study. The details of the models including calibration and parameterization can be found in Mahmood et al. (2023). The physics-based model was calibrated on the current 4 clusters and some other catchments without any holdouts. Moreover, the calibration time was longer than the verification period (Table A1) that surely impacts the MIKE-SHE model performance positively.

Table 3 Hyperparameter tunning of XGBoost

Hyper parameters	Description	Tested values	Optimized value Model1	Optimized value Model2	Optimized value Model3	Optimized value Model4
Reg_alpha	L1 regularization term on weights	0.0001, 0.001, 0.01, 0.1, 1, 10,100	1	1	1	1
Reg_lamda	L2 regularization term on weights	0.0001,0.001, 0.01, 0.1, 1, 10,100	1	1	1	1
Gamma	specifies the minimum loss reduction required to make a split.	0,0.1,0.2,0.3,0.4,0.5	0.4	0	0	0
Learning rate	specifies how fast a model learns	0.001, 0.008, 0.005, 0.01, 0.08, 0.05,0.1, 0.8,0.5,1	0.05	0.01	0.008	0.001
Max depth	maximum depth of a tree	3,6,9,12,15,18,21	15	12	13	15
Colsample_bytree	the fraction of columns to be randomly samples for each tree	0.3,0.5,0.7,0.9,1.1,1.3	0.9	0.7	0.7	0.7
n_estimators	the number of runs XGBoost will try to learn	100,500,1000,2000,4000,6000, 8000	4000	2000	8000	8000
Loss_function	function used to calculate the difference between input and output	RMSE, MAE, RMSLE	RMSLE	RMSLE	RMSE	RMSE
High drain flow data replication	addition of highest drain flows (1%) of data in training dataset one time or multiple times	No addition,1time, 4 times, 6 times	6 times	No addition	6 times	6 times

3. Results and discussion

3.1. Model cross-validation

The cross-validation scatterplots between predicted and observed drain flow of the eight ML models and the MIKE-SHE model are shown in Figure 4. The predicted drain flows from all five folds of cross-validation are combined for each model. For XGBoost the cross validation R^2 ranges between 0.7 and 0.93; the lower for Lolland cluster-model 2 and higher for Norsminde cluster-model 3 and Gyldenholm cluster-model 4. For CNN, the cross-validation R^2 ranged from 0.51 to 0.96, where Norsminde cluster-model 3 showed the highest R^2 value and Gyldenholm cluster-model 4 showed the lowest R^2 value. The scatter plot of the MIKE-SHE model depict the predicted and observed drain flow of the calibration period (see calibration period in Table A1). The R^2 value between the observed and predicted drain flow value was 0.55 (Figure 4).

3.2. Model verification

As expected, verification showed lower performance than the cross-validation (Figure 5). XGBoost models showed higher performance than the CNN models. Gyldenholm cluster-Model 4 performed highest with R^2 values between 0.50 and 0.53 for both CNN and XGBoost while Norsminde cluster-model 3 showed highest performance in XGBoost and lowest performance in CNN. Lillebaek cluster-model 1 and Lolland cluster-model 2 performed similar with R^2 value around 0.3 in XGBoost and CNN both. The MIKE-SHE model verification results also showed an R^2 value of 0.34 (see verification time period in Table A1, Figure 5). In general, none of the models (ML or physics-based) showed very strong performance. However, the XGBOOST results are as good or better than those of the MIKE-SHE model.

Examining the results more closely, we observed that extremely high drain flow values were not predicted accurately by any of the models (Figure 5). Figure 6 shows a hydrograph for drain site LK-1. It is not representative of all drain sites, but it does highlight the common issue of simulating peak flows by the ML and physics-based model. The hydrograph contains predicted and observed drain flows for all the models along with their residuals. The hydrographs make it clear that the general time series of drain flow is well represented, but the peak values are mostly underestimated or sometimes overestimated. This is further supported by the negative PBIAS of -47.2 and -37 for XGBoost and CNN respectively, while a positive PBIAS of 21 for MIKE-SHE (Figure 6).

https://doi.org/10.5194/egusphere-2023-1872
Preprint. Discussion started: 31 August 2023

Table 4 Structure and hyperparameters tunning of CNN

Hyperparameters	Description	Tested values	Optimized value Model1	Optimized value Model2	Optimized value Model3	Optimized value Model4
Optimizer	It updates the network parameters to optimize the loss function	Adam	Adam	Adam	Adam	Adam
Epochs	no. of iteration of training of train dataset	100	100	100	100	100
Learning rate	specifies how fast a model learns	0.0001, 0.001,0.01,0.1	0.0001	0.0001	0.0001	0.0001
Batch size	no. of training steps	10,15,20,25,30,40,50,60,80	20	20	60	10
Dropout	percentage of input unit dropped during training	0,0.1,0.3,0.5	0	0	0	0
First conv 1d (filter, kernel size)	filter generates multiple feature maps from the given features (increasing dimensionality) kernel size specifies the no. of previous time steps it takes at a time	Filter=32,64,128 Kernel size=(3,5,10)	(32,7)	(128,7)	(128,5)	(32,7)
Second conv 1d (filter, kernel size)		Filter=64,128,256 Kernel size=3,2	(64,4)	(265,4)	(256,3)	(64,4)
Pooling 1	Down-sample the feature maps	2	2	2	-	2
Third conv 1d (filter, kernel size)	filter generates multiple feature maps from the given features (increasing dimensionality) kernel size decides the no. of previous time steps it takes at a time	Filter=128, 256, 512 Kernel size=3,2	(128,2)	(512,2)	(512,3)	(128,2)
Pooling 2	downsample the feature maps	2	2	2	2	2
First dense layer	specifies no. of learnable parameters or weights	128,64,32,16	32	64	128	64
Second dense layer		64,32,16,8,1	8	32	1	32
Third dense layer		32,16,8,1	1	8	-	8
Fourth dense layer		1	-	1	-	1
Loss function	function used to calculate the difference between input and output	RMSE, RMSLE, MAE	RMSE	RMSE	RMSE	RMSE
High drain flow data replication	addition of highest drain flows (1%) of data in training dataset one time or multiple times	No addition, 1 time, 4 times, 6 times	no	no	1 time	no
Shuffle	True, False	True, False	True	False	True	False

https://doi.org/10.5194/egusphere-2023-1872
Preprint. Discussion started: 31 August 2023

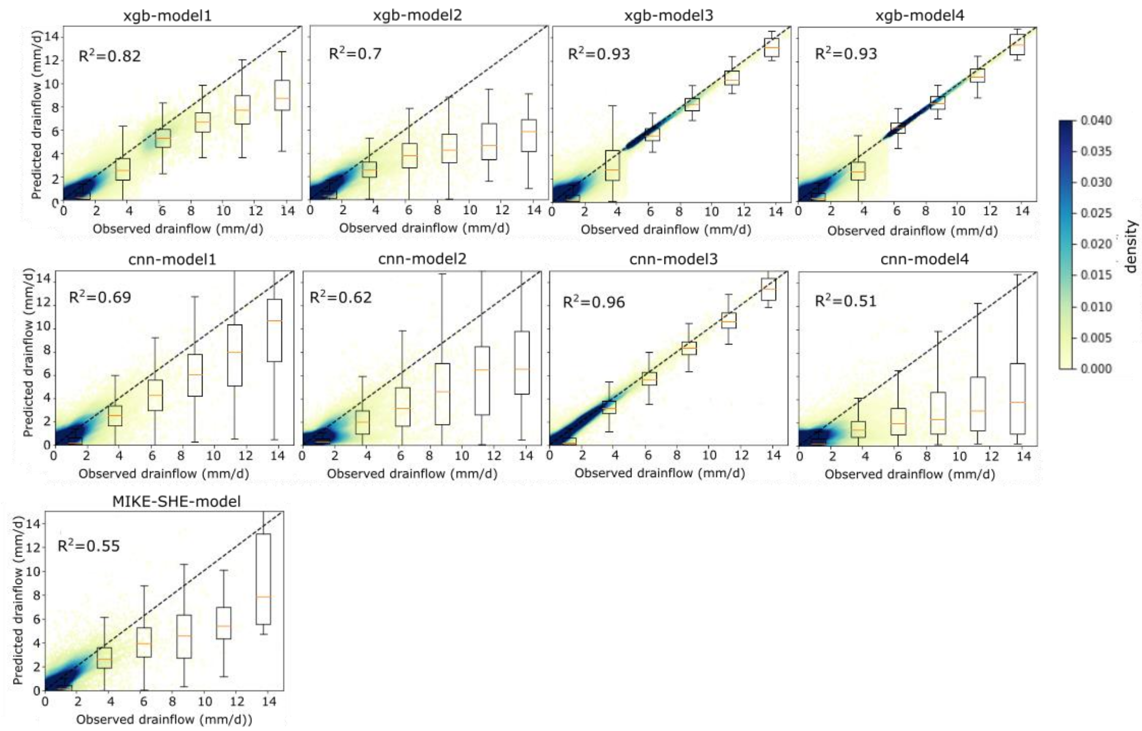

Figure 4 Five-fold cross-validation results for 4 models of XGBoost and 4 models of CNN. Density and boxplot of observed drain flow vs predicted drain flow.

https://doi.org/10.5194/egusphere-2023-1872
Preprint. Discussion started: 31 August 2023

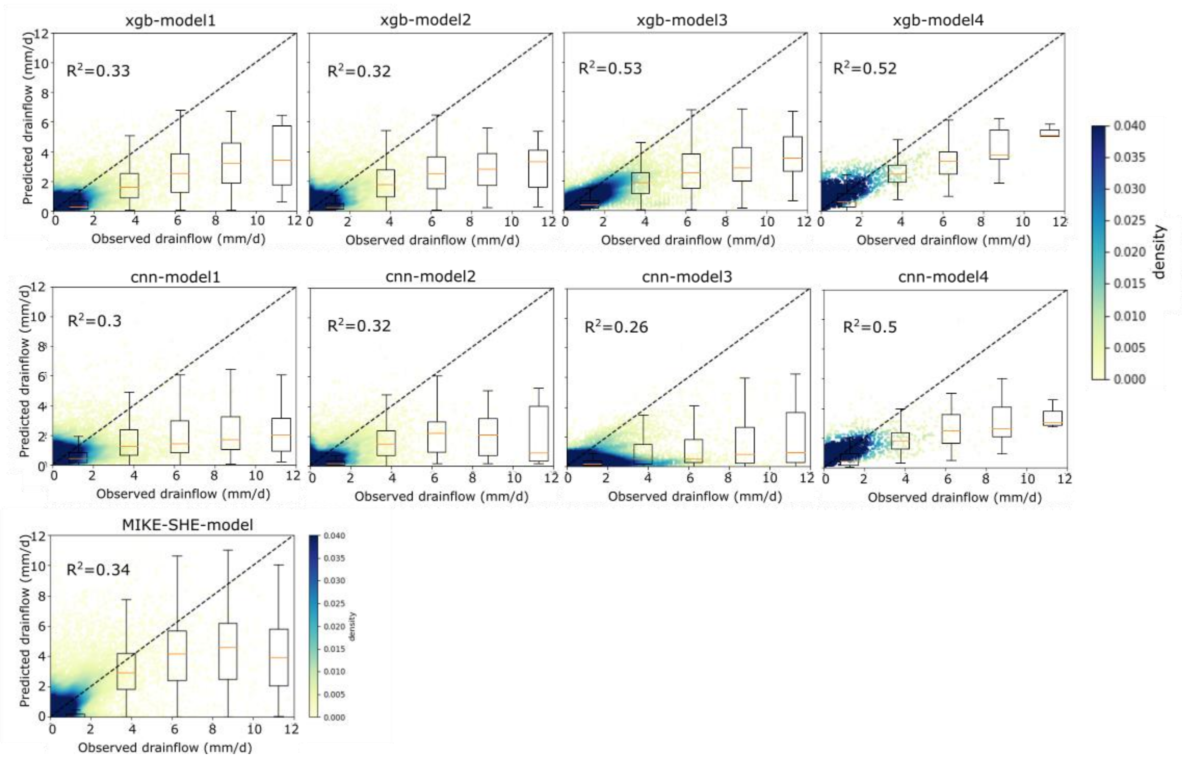

Figure 5 Verification results for 4 models of XGBoost and 4 models of CNN. Density and boxplot of observed drain flow vs predicted drain flow. Each scatterplot shows the mean of verification results obtained from 5-fold sub-models

https://doi.org/10.5194/egusphere-2023-1872
Preprint. Discussion started: 31 August 2023

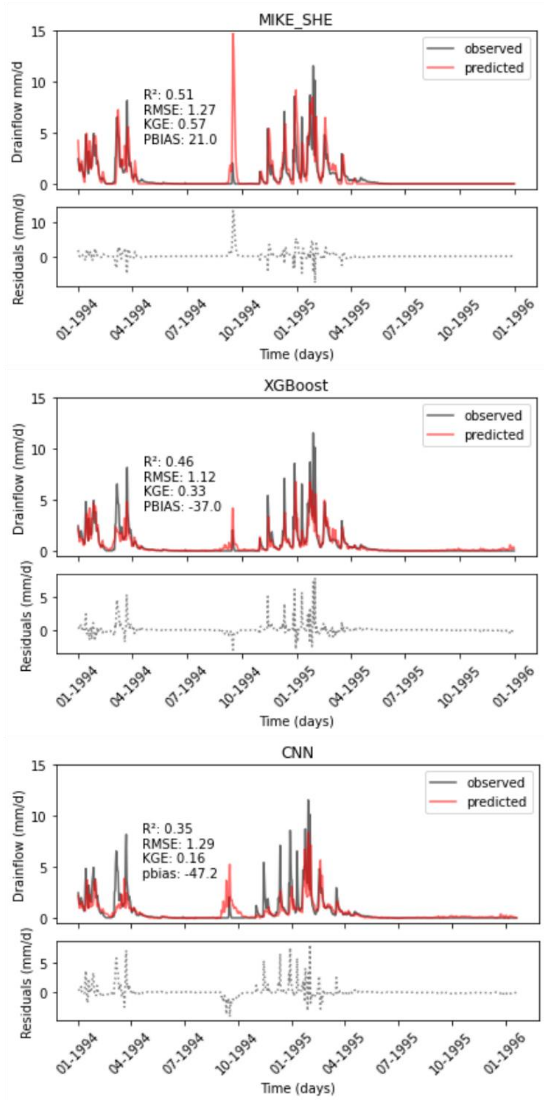

Figure 6 Hydrograph example of predicted and observed drain flow with additional plot of residuals between predicted and observed drain flow.

3.3. Comparison between MIKE-SHE, CNN and XGBoost performance

The KGE and PBIAS plots of cross-validation results across all drain sites are shown in Figure 7. Among the three models, XGBoost depicted the highest mean KGE of 0.72 across drain sites while CNN and MIKE-SHE showed mean KGE values of 0.61 and 0.59, respectively. In terms of PBIAS, MIKE-SHE (-3.31) performed better than XGBoost and CNN (5.82 and 7.79. respectively).

https://doi.org/10.5194/egusphere-2023-1872
Preprint. Discussion started: 31 August 2023

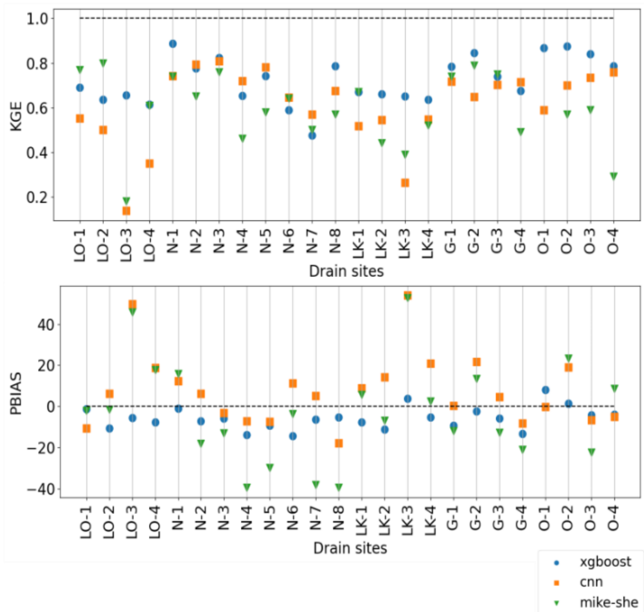

Figure 7 Comparison of average catchment performances of 4 CNN and 4 XGBoost models with MIKE-SHE model for cross-validation. The dotted line shows the highest achievable KGE and PBIAS

The KGE and PBIAS plots for the verification results across all drain sites are shown in Figure 8. Among the three models, MIKE-SHE showed the highest mean KGE of 0.42 across drain sites, followed by XGBoost with a mean KGE value of 0.35 across drain sites. CNN did not perform well with an average KGE value of 0.12 and a PBIAS value of -15.95 across drain sites. In terms of PBIAS, the XGBoost models performed highest with the lowest mean PBIAS value of 2.65, followed by MIKE-SHE model with PBIAS of 3.05. Among the drain sites, the Gyldenhom sites (G) performed best.

The comparison between ML models and the physics-based MIKE-SHE models revealed that MIKE-SHE performed similarly to XGBoost the better performing ML model. However, it is important to acknowledge that the fairness of the comparison between MIKE-SHE and machine learning models is compromised due to the differing techniques employed. In our machine learning models, we utilized the "Leave one cluster out" technique, whereas this approach was not employed in the MIKE-SHE model. The MIKE-SHE model, on the other hand, utilized most of the drain flow daily data from all drain sites for calibration (equivalent to cross-validation in machine learning), reserving approximately 10% of the drain flow data for verification (Table A1). This discrepancy in techniques has a significant impact on XGBoost and CNN model performance, as demonstrated

https://doi.org/10.5194/egusphere-2023-1872
Preprint. Discussion started: 31 August 2023

by Motarjemi et al. (2021) who employed the "leave spatially close sites out" technique and observed a drastic decrease in the predictive performance of various machine learning methods for annual drain flow prediction.

Furthermore, considering the amount of field data required and the time-intensive computational efforts involved in building a MIKE-SHE model, machine learning models emerge as a clear choice. Gumiere et al. (2020) compared the period to find appropriate solutions for physics-based hydrological model and ML model. They stated that the machine learning model improves performance and finds acceptable solutions in shorter lead times (3hrs) compared to the physics-based models (20hrs) due to autoregressive ability of ML models (Gumiere et al., 2020). They also highlighted that physics-based models require additional efforts to accurately depict the boundary conditions and parameter heterogeneity that are comparatively less in ML models. Firstly, it is hard to gather data on parameter heterogeneity and accurate boundary conditions. Secondly, it makes the model computationally more expensive. Machine learning models also need large datasets and catchment properties, but the current proposed model does not crucially require difficult to obtain field heterogeneity parameters and boundary conditions. Therefore, the flexibility and efficiency offered by machine learning models make them an advantageous option for drain flow prediction in comparison to MIKE-SHE.

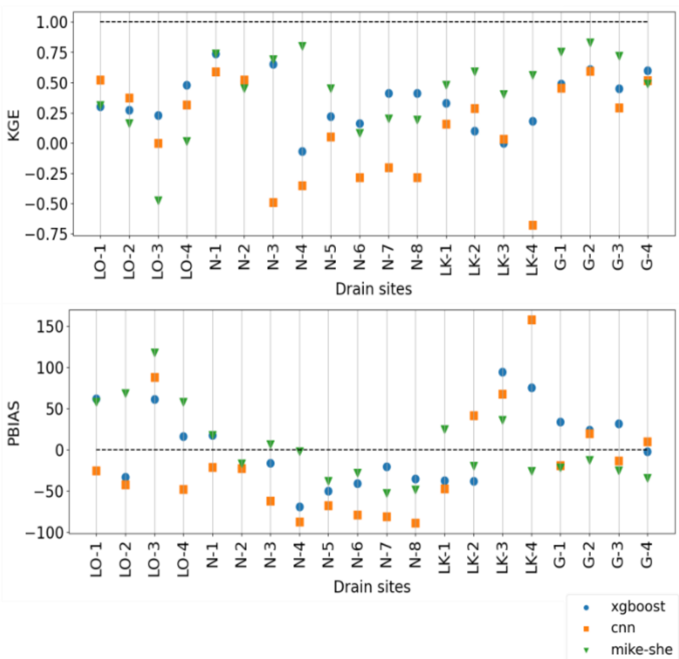

Figure 8 Comparison of average catchment performances of 4 CNN and 4 XGBoost models with MIKE-SHE model for verification. The dotted line shows the highest achievable KGE and PBIAS.

https://doi.org/10.5194/egusphere-2023-1872
Preprint. Discussion started: 31 August 2023

3.4. Feature importance

The average feature importance across 4 models of XGBoost models and CNN models is shown in Figure 9 for the verification dataset. We observed that the climate features played an important role in predicting the drain flows while static topographical and geological variables were less important. The most important climatic features in both XGBoost and CNN were mean precipitation, hydrological day of the year, mean evapotranspiration, mean temperature, and mean precipitation of the previous week. Among the topographical features, mean TWI of catchment was the eighth most important feature while the standard deviation in TWI of the 300 m buffer outside the catchment was the fifth most important feature. The geological features were considered least important among both XGBoost and CNN models.

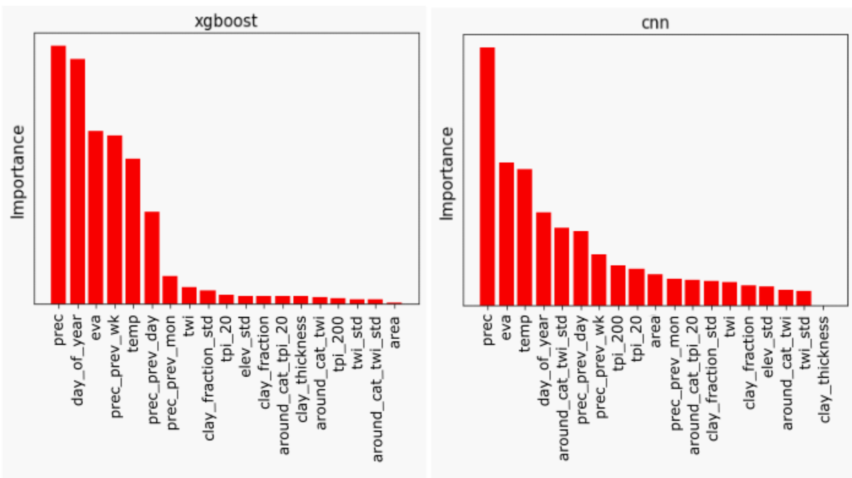

Figure 9 Average feature importance across 4 models for CNN and XGBoost

The importance of the most significant climatic features in the CNN and XGBoost model is evident. Although machine learning models don’t take physics into account, the identified important features were consistent with the physics behind it. Precipitation emerged as the primary driver of drain flow, while temperature and evapotranspiration exhibited inverse correlations (Figure 10). As temperature increases, evapotranspiration intensifies, limiting the water available for drain flows. The hydrological day of the year ranked fourth in importance in the CNN model and second in the XGBoost model. This pointed to the capturing of seasonal variations in drain flows. Specifically, the hydrological day of the year reflected higher drain flows during winter months and minimal drain flows during summer months, aligning with the expected hydrological patterns (Figure

https://doi.org/10.5194/egusphere-2023-1872
Preprint. Discussion started: 31 August 2023

10). Overall, the CNN and XGBoost model effectively captured these relationships, shedding light on the interplay between climatic factors and hydrological processes.

In XGBoost and CNN, the precipitation in previous day, week and month were also found to be important factors in predicting drain flow. This in terms of physics indicated that constantly high precipitation occurred on the previous day, week or month could be a good indicator of high groundwater levels (above drain depth) and soil saturation that drives the drain flows. Conversely, no precipitation event in the past day, week or month can make groundwater level lower (below drain level) and soil dry. This could lead to low to no drain flows. However, the influence of topographical features neither depicted a clear correlation nor they were found among the important features in ML models (Figure 10g & f).

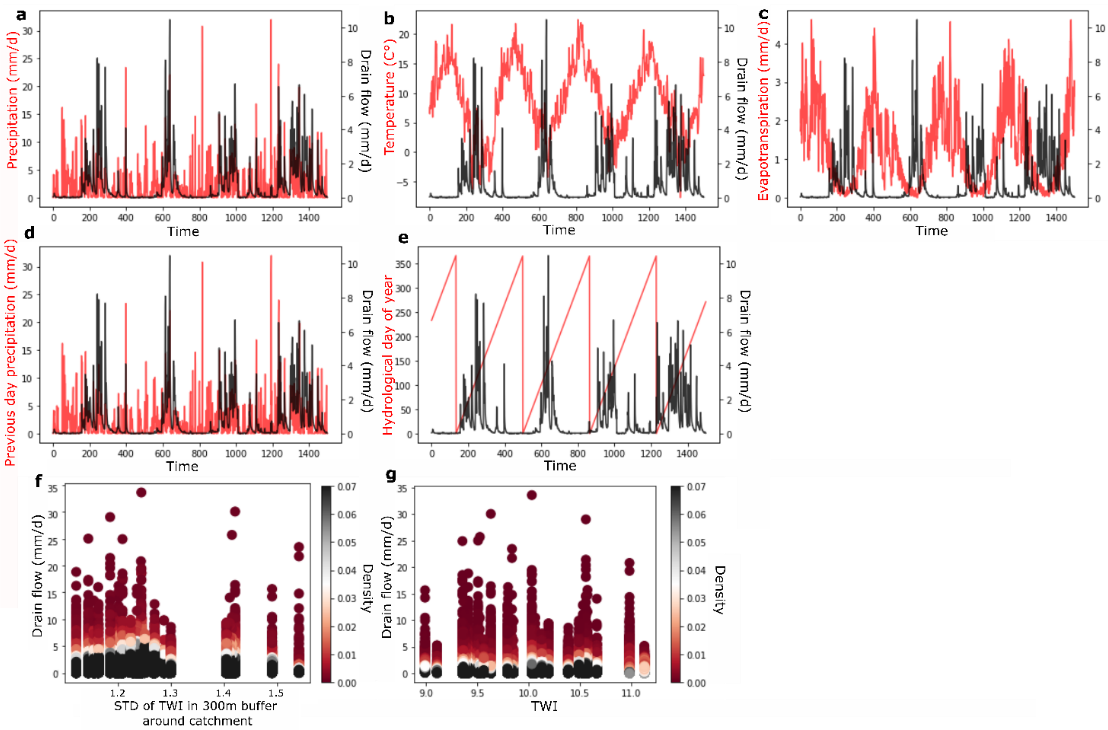

Figure 10 Plots with drain flow and important features of CNN and XGBoost.

3.5. Model performance of CNN and XGBoost

The comparison of CNN and XGBoost model showed that the model performance of XGBoost was better than CNN. This might be because XGBoost can work better with relatively small datasets while neural networks such as CNN require larger dataset for better predictions (Gauch et al., 2021). Similar to our findings, Motarjemi et al.

https://doi.org/10.5194/egusphere-2023-1872
Preprint. Discussion started: 31 August 2023

(2021) also found that annual drain flows were better predicted using random forest and cubist machine learning methods than neural networks.

Despite achieving superior performance compared to CNN in cross-validation and verification datasets (Figure 7, Figure 8), both XGBoost and CNN models failed to accurately predict high drain flow events during the verification phase (depicted in Figure 6). In the case of XGBoost, the overall model performance across different clusters, encompassing all four models, exhibited an intermediate to weak performance range, with R^2 values ranging from 0.32 to 0.53. However, our analysis revealed a strong relationship between the model performance and the normalized drain flow at 10% exceedance probability observed across all drain sites. Figure 11a shows the normalized drain flow at 10% exceedance probability across all drain sites. The KGE value increases with the normalized drain flow at 10% exceedance probability. This is depicted by the high Spearman correlations of 0.72 and 0.74 between KGE values and normalized drain flow at 10% exceedance probability for both XGBoost cross-validation and verification performances respectively (Figure 11b). Cross-validation KGE of CNN also showed a high spearman correlation of 0.68 with normalized drain flow at 10% exceedance probability, however the verification KGE of CNN showed a weak correlation.

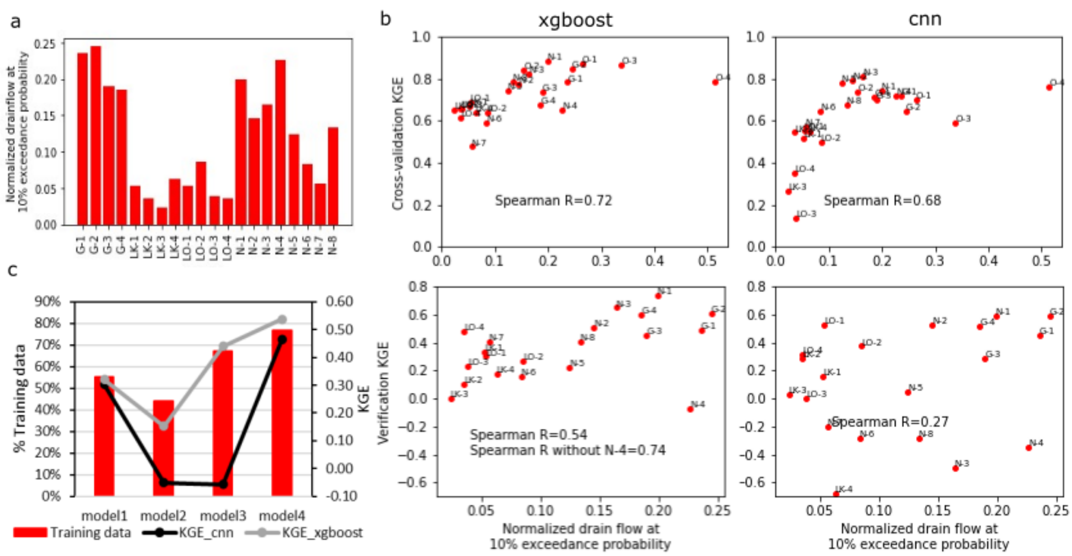

Figure 11 Model performance linked with normalized drain flow 10% exceedance probability and training dataset; Normalized drain flow at 10% exceedance probability across all drain sites (a); Normalized drain flow at 10% exceedance probability vs cross-validation KGE and verification KGE for CNN and XGBoost (b); trend of training data % compared to model performances of XGBoost and CNN (c).

https://doi.org/10.5194/egusphere-2023-1872
Preprint. Discussion started: 31 August 2023

The 10% exceedance probability of normalized drain flow indicated a peakier drain flow behavior, which is more difficult to simulate well. Even though we attempted to increase the weightage of peaks by doubling high drain flow values in the training dataset but it could be that there was still relatively less representation of peak flows in data so ML model could not learn it better. The accuracy of the collected field data is another influential factor that can explain the inability of models to simulate peak flow. Previous studies have indicated that even physics-based models like MIKE-SHE struggled to accurately simulate drain flows from the Lolland and Lillebaek drain sites (Hansen et al., 2013; Mahmood et al., 2023). This suggests amongst others, potential issues with the data quality at these sites especially in drain flow observations. Additionally, the precipitation data utilized in our machine learning models were collected from a station-based product interpolated to a 10 km grid, which may not effectively capture the local precipitation patterns at the field scale of the drain sites ranging from 1 to 100 ha.

Generally, the high drain flows could either be linked to high precipitation or high lateral flows. Figure 10 provides evidence that drain flows sometimes exceeded the precipitation, particularly during winter. Although the machine learning model incorporated climatic variables (precipitation), but the peaky response from the drain flows was not fully learned by the ML models indicating that we missed predictor variables that can represent upward lateral flows towards the tile drains. On the contrary, distributed physics-based models like MIKE-SHE can easily capture the influence of lateral flows on tile drain flows. Moreover, while certain topographical aspects like TWI and TPI were included as feature variables, the model had limited representation of geological factors such as hydraulic conductivity and other pedogeological variables that can influence the peak flows.

Our analysis also revealed a discernible pattern between the percentage of training data and the resulting model performance for both CNN and XGBoost models, as depicted in Figure 11c. The trend demonstrated a positive correlation, indicating that higher percentages of training data corresponded to higher KGE values. This observation aligns with the findings of Joseph (2022) who investigated the impact of data split on model performance and concluded that a lower ratio between training and testing data leads to reduced model performance. In our study, the varying lengths of time series data across different clusters significantly influenced the model performance, contributing to the observed pattern.

The current performance of ML models leaves room for uncertainty regarding their applicability across different time periods and geographical locations. Notably, there is currently no existing daily tile-drain flow prediction model available that is developed for multiple drain sites. The scarcity of such models could be attributed to the requirement for a more comprehensive dataset encompassing temporal and spatial variations in drain flow

https://doi.org/10.5194/egusphere-2023-1872
Preprint. Discussion started: 31 August 2023

behaviors. However, we maintain a positive outlook on the potential improvement of model performance through an increased collection of drain flow data from diverse drain sites. Our optimism is supported by the findings of Kratzert et al. (2019) who effectively predicted a regional rainfall runoff model using neural network technique. Their research demonstrated the feasibility of utilizing a single machine learning model to capture both regional and local-scale trends in the runoff model by incorporating streamflow data from 531 catchments across the United States, thereby accounting for the temporal and spatial diversity within their training dataset.

Overall, the study developed two ML models to investigate whether these models could be used to predict drain flow over space and time in Denmark. However, none of the ML models showed high performance. The Gyldenholm cluster model did perform satisfactory based on its R2 value, but the other three models had weak performance. This shows that current available data is insufficient to develop a transferable model, especially in space. More training data will be required especially in terms of number of drain sites from different clusters to include drain flows from various topographical and geological features. Gauch et al. (2021) also suggested the same for stream flow prediction model. We believe this because drain flow time series data performs well when some of the drain sites from each cluster are involved in training dataset (Motarjemi et al., 2021). Krizhevsky et al. (2017) also proved that increase in training dataset improves the ML model performance. In addition to that, physics guided machine learning can also be an effective and efficient solution to proceed in ML based drain flow model predictions. Incorporation of parameters such as groundwater heads or regional lateral fluxes generated from the physics-based model could improve the ML model performance. In addition to that, physics guided machine learning can also be an effective and efficient solution to proceed in ML based drain flow model predictions. This new approach can be advantageous by incorporating the existing physical understanding of drain flow models with the strengths of data driven model. The output of the physics guided ML models can be more generalizable and transferable (Willard et al., 2020; Yang et al., 2019).

Conclusion

This study explored the potential of two ML models; XGBoost and CNN and later compared its performance to the existing physics-based MIKE-SHE models. The results indicated that XGBoost models consistently outperform CNN models, as demonstrated by relatively higher accuracy in both cross-validation and verification stages. Overall, the XGBoost performed decently to predict timing and lower flows but did not perform well in peak flows.

The study also showed that the XGBoost model performed as well as the physics-based MIKE-SHE models. We found that although ML models do not include the physical processes explicitly, the ML model can capture missing processes unlike physics-based models because of its ability to learn complex nonlinear patterns from the data. This is because ML models do not need physical processes equation therefore, allowing them to extract insights directly from the data. Additionally, ML models are easier to build and train than the physics-based models and can provide valuable insight into the controlling features of the drain flows. We believe that ML models have the potential to replace the traditional physics-based models once sufficient data is available to make a more transferable model.

Acknowledgements

This study was part of T-Rex (Terrænnær redox og retentions-kortlægning til differentieret målrettet virkemiddelindsats indenfor ID15 oplande), funded by the Danish GUDP (Grønt Udviklings- og Demonstrationsprogram), project number 34009–18-1453 and WATEC – Aarhus University Center for Water Technology in Denmark. Authors would also like to thank Muhammad Rizwan Asif for productive discussions and for sharing his knowledge of Machine learning. I would also like to acknowledge that a few sentences in manuscript were improved for sentence structure and grammar using AI tool.

Author contribution

Hafsa Mahmood: Conceptualization, Methodology, Software, Formal analysis, Investigation, Writing-original draft, Visualization, Validation**, Ty P. A. Ferré:** Conceptualization, Methodology, supervision, review and editing, **Raphael M. J. Schneider:** Conceptualization, Methodology, supervision, review and editing, **Simon Stisen:** Investigation, Methodology, Review & editing, Supervision, **Rasmus R. Frederiksen:** Review & editing, Supervision, **Anders V. Christiansen:** Review & editing, Supervision.

Declaration of interest

The authors declare that they have no conflict of interest.

Data Availability

Data can be available on request.

https://doi.org/10.5194/egusphere-2023-1872
Preprint. Discussion started: 31 August 2023

Appendix A

Table A1 MIKE-SHE model calibration and verification period

Cluster	*Drain sites*	*MIKE-SHE calibration period start*	*MIKE-SHE calibration period end*	*MIKE-SHE verification period start*	*MIKE-SHE verification period end*
"other"	**Fillerup**	01-01-2014	12-31-2016	01-01-2017	29-05-2017
	Ulvsborg	01-01-2016	01-05-2018	-	-
	Vadum	01-01-2014	12-31-2016	-	-
Norsmlnde	**Norsminde1**	01-01-2014	12-31-2016	01-01-2017	08-06-2017
	Norsminde2	01-01-2014	12-31-2016	01-01-2017	18-06-2017
	Norsminde3	01-01-2014	12-31-2016	01-01-2017	27-06-2017
	Norsminde4	01-01-2014	12-31-2016	01-01-2017	27-06-2017
	Norsminde5	01-01-2014	12-31-2016	01-01-2017	08-06-2017
	Norsminde6	01-01-2014	12-31-2016	01-01-2017	06-02-2018
	Norsminde7	01-01-2014	12-31-2016	01-01-2017	31-12-2018
	Norsminde8	01-01-2014	12-31-2016	01-01-2017	13-12-2018
Lillebæk	**Lillebaek1**	01-01-1997	12-31-1999	01-01-1994	31-12-1995
	Lillebaek2	01-01-1997	12-31-1999	01-01-1994	31-12-1995
	Lillebaek3	01-01-1997	12-31-1999	01-01-1994	31-12-1995
	Lillebaek4	01-01-1997	12-31-1999	01-01-1994	31-12-1995
Lolland	**Lolland1**	01-01-2014	12-31-2016	01-01-2000	01-01-2005
	Lolland2	01-01-2014	12-31-2016	01-01-2000	01-01-2005
	Lolland3	01-01-2014	12-31-2016	01-01-2000	01-01-2005
	Lolland4	01-01-2014	12-31-2016	01-01-2000	01-01-2005
Gyldenholm	**Gyldenholm1**	01-01-2016	12-31-2017	01-01-2018	09-05-2018
	Gyldenholm2	01-01-2016	12-31-2017	01-01-2018	09-05-2018
	Gyldenholm3	01-01-2016	12-31-2017	01-01-2018	09-05-2018
	Gyldenholm4	01-01-2016	12-31-2017	01-01-2018	09-05-2018

References

Abadi, M., Barham, P., Chen, J. M., Chen, Z. F., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., Kudlur, M., Levenberg, J., Monga, R., Moore, S., Murray, D. G., Steiner, B., Tucker, P., Vasudevan, V., Warden, P., Wicke, M., Yu, Y., & Zheng, X. Q. (2016). TensorFlow: A system for large-scale machine learning. *Proceedings of Osdi'16: 12th Usenix Symposium on Operating Systems Design and Implementation*, 265-283. <Go to ISI>://WOS:000569062400017

Adhikari, K., Kheir, R. B., Greve, M. B., Bocher, P. K., Malone, B. P., Minasny, B., McBratney, A. B., & Greve, M. H. (2013, May). High-Resolution 3-D Mapping of Soil Texture in Denmark. *Soil Science Society of America Journal, 77*(3), 860-876. https://doi.org/10.2136/sssaj2012.0275

Bai, S., Kolter, J. Z., & Koltun, V. (2018). An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling. *ArXiv. /abs/1803.01271*.

Basha, E. A., Ravela, S., & Rus, D. (2008). Model-Based Monitoring for Early Warning Flood Detection. *Sensys'08: Proceedings of the 6th Acm Conference on Embedded Networked Sensor Systems*, 295-308. <Go to ISI>://WOS:000267822600022

Bechtold, M., Tiemeyer, B., Laggner, A., Leppelt, T., Frahm, E., & Belting, S. (2014). Large-scale regionalization of water table depth in peatlands optimized for greenhouse gas emission upscaling. *Hydrology and Earth System Sciences, 18*(9), 3319-3339. https://doi.org/10.5194/hess-18-3319-2014

Beven, K. (1989, Jan 30). Changing Ideas in Hydrology - the Case of Physically-Based Models. *Journal of Hydrology, 105*(1-2), 157-172. https://doi.org/Doi 10.1016/0022-1694(89)90101-7

Bjerre, E., Fienen, M. N., Schneider, R., Koch, J., & Hojberg, A. L. (2022, Sep). Assessing spatial transferability of a random forest metamodel for predicting drainage fraction. *Journal of Hydrology, 612*. https://doi.org/ARTN 128177

10.1016/j.jhydrol.2022.128177

Chen, T. Q., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Kdd'16: Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining*, 785-794. https://doi.org/10.1145/2939672.2939785

Craft, K. J., Helmers, M. J., Malone, R. W., Pederson, C. H., & Schott, L. R. (2018). Effects of Subsurface Drainage Systems on Water and Nitrogen Footprints Simulated with Rzwqm2. *Transactions of the Asabe, 61*(1), 245-261. https://doi.org/10.13031/trans.12300

Dai, Z. H., Zhang, M., Nedjah, N., Xu, D., & Ye, F. (2023, Feb). A Hydrological Data Prediction Model Based on LSTM with Attention Mechanism. *Water, 15*(4). https://doi.org/ARTN 670

10.3390/w15040670

De Schepper, G., Therrien, R., Refsgaard, J. C., He, X., Kjaergaard, C., & Iversen, B. V. (2017, May). Simulating seasonal variations of tile drainage discharge in an agricultural catchment. *Water Resources Research, 53*(5), 3896-3920. <Go to ISI>://WOS:000403712100024

DMI, D. M. D. (2023).

EPA, D. (2020). *FOHM—Fælles Offentlig Hydrologisk Model* https://mst.dk/natur-vand/vand-i-hverdagen/grundvand/grundvandskortlaegning/kortlaegning-2016-2020/fohm-faelles-offentlig-hydrologisk-model/

Erickson, M. L., Elliott, S. M., Brown, C. J., Stackelberg, P. E., Ransom, K. M., & Reddy, J. E. (2021, Apr). Machine Learning Predicted Redox Conditions in the Glacial Aquifer System, Northern Continental United States. *Water Resources Research, 57*(4). https://doi.org/ARTN e2020WR02820710.1029/2020WR028207

Frederiksen, R. R., Larsen, S. E., Mathiesen, G. B., & Kronvang, B. (2023). Development and application of a parsimonious statistical model to predict tile flow in minerogenic soils. *Agricultural Water Management*.

Gauch, M., Mai, J., & Lin, J. (2021). The proper care and feeding of CAMELS: How limited training data affects streamflow prediction. *Environmental Modelling and Software, 135, 104 926*. https://doi.org/https://doi.org/10.1016/j.envsoft.2020.104926

Golmohammadi, G., Rudra, R. P., Parkin, G. W., Kulasekera, P. B., Macrae, M., & Goel, P. K. (2021, Mar). Assessment of Impacts of Climate Change on Tile Discharge and Nitrogen Yield Using the DRAINMOD Model. *Hydrology, 8*(1). https://doi.org/ARTN 1

10.3390/hydrology8010001

Gumiere, S. J., Camporese, M., Botto, A., Lafond, J. A., Paniconi, C., Gallichand, J., & Rousseau, A. N. (2020, Apr 21). Machine Learning vs. Physics-Based Modeling for Real-Time Irrigation Management. *Frontiers in Water, 2*. https://doi.org/ARTN 8

10.3389/frwa.2020.00008

Gupta, H. V., Kling, H., Yilmaz, K. K., & Martinez, G. F. (2009, Oct 20). Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. *Journal of Hydrology, 377*(1-2), 80-91. https://doi.org/10.1016/j.jhydrol.2009.08.003

Hansen, A. L., Refsgaard, J. C., Christensen, B. S. B., & Jensen, K. H. (2013). Importance of including small-scale tile drain discharge in the calibration of a coupled groundwater-surface water catchment model. *Water Resources Researc*. https://doi.org/ https://doi.org/10.1029/2011WR011783

Herath, H. M. V. V., Chadalawada, J., & Babovic, V. (2021). Hydrologically informed machine learning for rainfall–runoff modelling: towards distributed modelling, . *Hydrol. Earth Syst. Sci*(25), 4373–4401. https://doi.org/https://doi.org/10.5194/hess-25-4373-2021

Hoang, L., van Griensven, A., van der Keur, P., Refsgaard, J. C., Troldborg, L., Nilsson, B., & Mynett, A. (2014, Jan-Feb). Comparison and Evaluation of Model Structures for the Simulation of Pollution Fluxes in a Tile-Drained River Basin. *Journal of Environmental Quality, 43*(1), 86-99. https://doi.org/10.2134/jeq2011.0398

Jeantet, A., Thirel, G., Jeliazkov, A., Martin, P., & Tournebize, J. (2022, Jun 29). Effects of Climate Change on Hydrological Indicators of Subsurface Drainage for a Representative French Drainage Site. *Frontiers in Environmental Science, 10*. https://doi.org/ARTN 899226

10.3389/fenvs.2022.899226

Joseph, V. R. (2022). Optimal ratio of data split. *Statistical Analysis and Data Mining The ASA Data Sci Journal*.

https://doi.org/10.5194/egusphere-2023-1872
Preprint. Discussion started: 31 August 2023

Koch, J., Berger, H., Henriksen, H. J., & Sonnenborg, T. O. (2019, Nov 15). Modelling of the shallow water table at high spatial resolution using random forests. *Hydrology and Earth System Sciences, 23*(11), 4603-4619. https://doi.org/10.5194/hess-23-4603-2019

Koch, J., Gotfredsen, J., Schneider, R., Troldborg, L., Stisen, S., & Henriksen, H. J. (2021, Sep 1). High Resolution Water Table Modeling of the Shallow Groundwater Using a Knowledge-Guided Gradient Boosting Decision Tree Model. *Frontiers in Water, 3*. https://doi.org/ARTN 701726

10.3389/frwa.2021.701726

Koch, J., & Schneider, R. (2022). Long short-term memory networks enhance rainfall-runoff modelling at the national scale of Denmark. *Geus Bulletin, 49*. https://doi.org/ARTN 8292

10.34194/geusb.v49.8292

Kratzert, F., Klotz, D., Shalev, G., Klambauer, G., Hochreiter, S., & Nearing, G. (2019, Dec 17). Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets. *Hydrology and Earth System Sciences, 23*(12), 5089-5110. <Go to ISI>://WOS:000503445500001

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017, Jun). ImageNet Classification with Deep Convolutional Neural Networks. *Communications of the Acm, 60*(6), 84-90. https://doi.org/10.1145/3065386

Kuzmanovski, V., Trajanov, A., Leprince, F., Dzeroski, S., & Debeljak, M. (2015, Feb 1). Modeling water outflow from tile-drained agricultural fields. *Science of the Total Environment, 505*, 390-401. https://doi.org/10.1016/j.scitotenv.2014.10.009

Lawrence, S., Giles, C. L., Tsoi, A. C., & Back, A. D. (1997, Jan). Face recognition: A convolutional neural-network approach. *Ieee Transactions on Neural Networks, 8*(1), 98-113. https://doi.org/Doi 10.1109/72.554195

Lees, T., Reece, S., Kratzert, F., Klotz, D., Gauch, M., De Bruijn, J., Sahu, R. K., Greve, P., Slater, L., & Dadson, S. J. (2022, Jun 20). Hydrological concept formation inside long short-term memory (LSTM) networks. *Hydrology and Earth System Sciences, 26*(12), 3079-3101. https://doi.org/10.5194/hess-26-3079-2022

Lewinson, E. (2020). *Python for Finance Cookbook: Over 50 recipes for applying modern Python libraries to financial data analysis*. Packt publishing.

Mahmood, H., Schneider, R. M., Frederiksen, R. R., Christiansen, A. V., & Stisen, S. (2023). Using jointly calibrated fine-scale drain models across Denmark to assess the influence of physical variables on spatial drain flow patterns. *Journal of Hydrology: Regional Studies, 46*(101353). https://doi.org/https://doi.org/10.1016/j.ejrh.2023.101353

Mirlas, V. (2009, May-Jun). Applying MODFLOW Model for Drainage Problem Solution: A Case Study from Jahir Irrigated Fields, Israel. *Journal of Irrigation and Drainage Engineering, 135*(3), 269-278. <Go to ISI>://WOS:000266202900001

Moller, A. B., Beucher, A., Iversen, B. V., & Greve, M. H. (2018, Jun 15). Predicting artificially drained areas by means of a selective model ensemble. *Geoderma, 320*, 30-42. https://doi.org/10.1016/j.geoderma.2018.01.018

Motarjemi, S. K., Moller, A. B., Plauborg, F., & Iversen, B. V. (2021, Aug). Predicting national-scale tile drainage discharge in Denmark using machine learning algorithms. *Journal of Hydrology-Regional Studies, 36*. https://doi.org/10.1016/j.ejrh.2021.100839

Muma, M., Gumiere, S. J., & Rousseau, A. N. (2014, Aug). Comprehensive analysis of the CATHY model sensitivity to soil hydrodynamic properties of a tile-drained, agricultural micro-watershed. *Hydrological Sciences Journal-Journal Des Sciences Hydrologiques, 59*(8), 1606-1623. https://doi.org/10.1080/02626667.2013.843778

Mushtaq, H., Akhtar, T., Hashmi, M. Z. R., & Masood, A. (2022). Hydrologic Interpretation of Machine Learning Models for 10-daily streamflow simulation in Climate sensitive Upper Indus Catchments. *Hydrology and Earth System Sciences*. https://doi.org/https://doi.org/10.5194/hess-2022-213

Northcott, W. J., Cooke, R. A., Walker, S. E., Mitchell, J. K., & Hirschi, M. C. (2001, Mar-Apr). Application of DRAINMOD-N to fields with irregular drainage systems. *Transactions of the Asae, 44*(2), 241-249. <Go to ISI>://WOS:000170723100008

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research 12*, 2825-2830

Scharling, M. (1999a). *Klimagrid Danmark - Nedbør, lufttemperatur og potentiel fordampning 20X20 & 40x40 km - Metodebeskrivelse*.

Scharling, M. (1999b). *Klimagrid Danmark Nedbør 10x10 km (ver. 2) - Metodebeskrivelse*.

Schneider, R., Koch, J., Troldborg, L., Henriksen, H. J., & Stisen, S. (2022, Nov 23). Machine-learning-based downscaling of modelled climate change impacts on groundwater table depth. *Hydrology and Earth System Sciences, 26*(22), 5859-5877. https://doi.org/10.5194/hess-26-5859-2022

Shen, C. (2018). A transdisciplinary review of deep learning research and its relevance for water resources scientists. *Water Resour. Res.*, 8558–8593. https://doi.org/10.1029/2018WR022643

Singh, J., Knapp, H. V., Arnold, J. G., & Demissie, M. (2005, Apr). Hydrological modeling of the iroquois river watershed using HSPF and SWAT. *Journal of the American Water Resources Association, 41*(2), 343-360. https://doi.org/DOI 10.1111/j.1752-1688.2005.tb03740.x

Tesoriero, A. J., Terziotti, S., & Abrams, D. B. (2015, Aug 18). Predicting Redox Conditions in Groundwater at a Regional Scale. *Environmental Science & Technology, 49*(16), 9657-9664. https://doi.org/10.1021/acs.est.5b01869

Willard, J., Jia, X., Xu, S., Steinbach, M., & Kumar, V. (2020). Integrating Physics-Based Modeling With Machine Learning: A Survey. *ACM Computing Surveys*.

Xu, W., Jiang, Y. N., Zhang, X. L., Li, Y., Zhang, R., & Fu, G. T. (2020, Dec). Using long short-term memory networks for river flow prediction. *Hydrology Research, 51*(6), 1358-1376. <Go to ISI>://WOS:000600213800009

Yang, T., Sun, F., Gentine, P., Liu, W., Wang, H., Yin, J., Du, M., & Liu, C. (2019). Evaluation and machine learning improvement of global hydrological model-based flood simulations. *Environmental Research Letter*, *19*. https://doi.org/DOI 10.1088/1748-9326/ab4d5e

Yao, G. L., Lei, T., & Zhong, J. D. (2019, Feb 1). A review of Convolutional-Neural-Network-based action recognition. *Pattern Recognition Letters*, *118*, 14-22. https://doi.org/10.1016/j.patrec.2018.05.018

Youssef, M. A., Liu, Y., Chescheir, G. M., Skaggs, R. W., & Negm, L. M. (2021). DRAINMOD modeling framework for simulating controlled drainage effect on lateral seepage from artificially drained fields. *Agricultural Water Management*. https://doi.org/https://doi.org/10.1016/j.agwat.2021.106944

Zia, H., Harris, N., Merrett, G., & Rivers, M. (2015, Oct). Predicting discharge using a low complexity machine learning model. *Computers and Electronics in Agriculture*, *118*, 350-360. <Go to ISI>://WOS:000364603500036