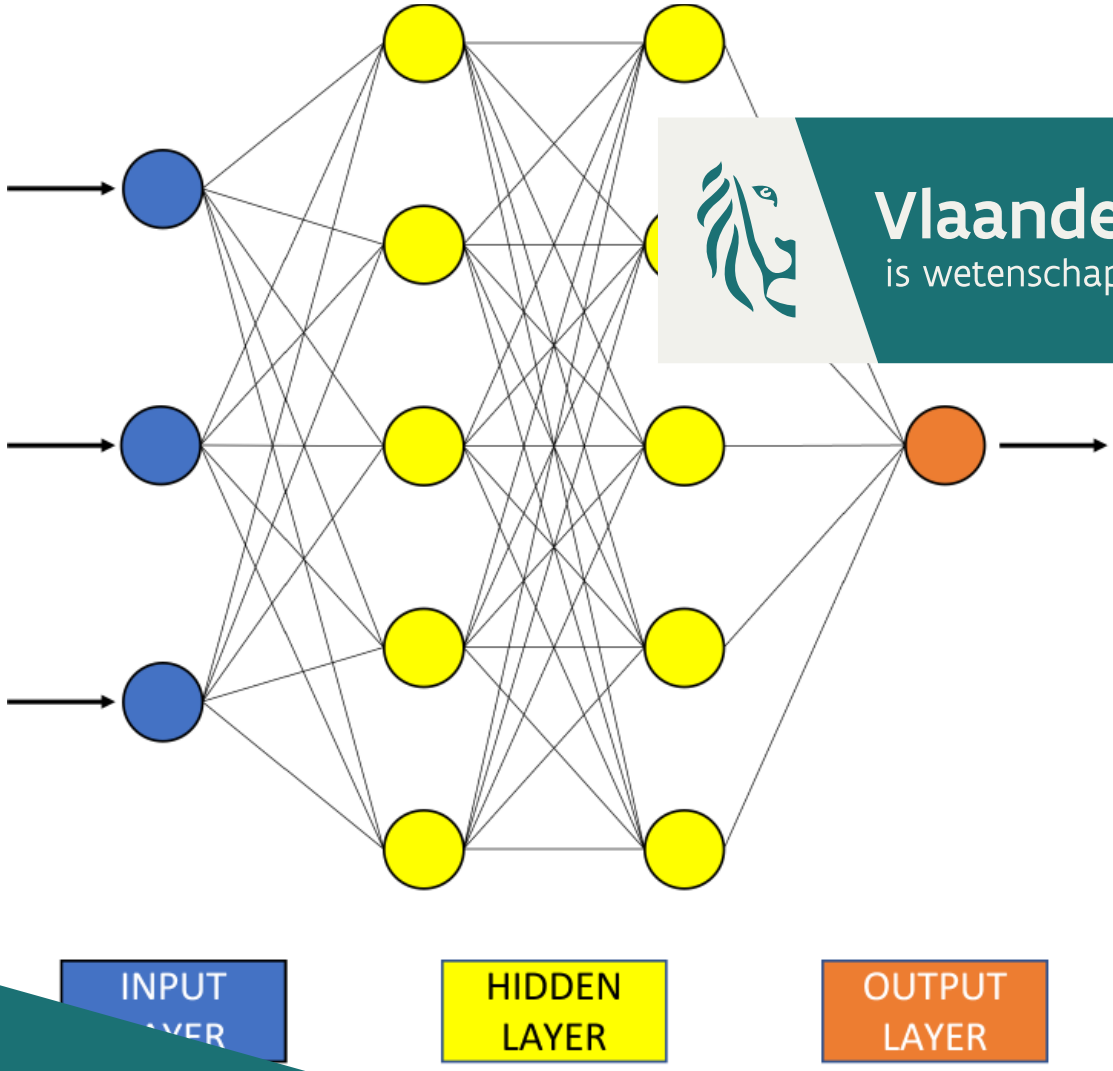


Vlaanderen
is wetenschap

17_070_1
WL rapporten

Exploreren van de mogelijkheden voor Big Data en Data-mining toepassingen

Eindrapport

DEPARTEMENT
MOBILITEIT &
OPENBARE
WERKEN

waterbouwkundiglaboratorium.be

Exploreren van de mogelijkheden voor Big Data en Data-mining toepassingen

Eindrapport

Nossent, J.

Waterbouwkundig Laboratorium
Vlaanderen is wetenschap

Cover figuur © Vlaamse overheid, Departement Mobiliteit en Openbare Werken, Waterbouwkundig Laboratorium

Juridische kennisgeving

Copyright en wijze van citeren

Deze publicatie dient als volgt geciteerd te worden:

Nossent, J. (2023). Exploreren van de mogelijkheden voor Big Data en Data-mining toepassingen: Eindrapport. Versie 1.0. WL Rapporten, 17_070_1. Waterbouwkundig Laboratorium: Antwerpen

Documentidentificatie

Opdrachtgever:	DMOW-WL	Ref.:	WL2023R17_070_1
Trefwoorden (3-5):	Big Data; Machine Learning; Hoogwatervoorspellingen; Kennisopbouw		
Kennisdomeinen:	Waterbeheer > 5. Statistiek > 5.3. Big Data & Data Mining		
Tekst (p.):		Bijlagen (p.):	/
Vertrouwelijk:	☒ Nee	☒ Online beschikbaar	

Auteur(s):	Nossent, J.

Controle

	Naam	Handtekening
Revisor(en):	Bertels, J.	Getekend door:Jonas Bertels (Signature) Getekend op:2023-01-31 16:30:52 +01:0 Reden:Ik keur dit document goed Jonas Bertels
Projectleider:	Nossent, J.	Getekend door:Jiri Nossent (Signature) Getekend op:2023-01-30 11:26:37 +01:0 Reden:Ik keur dit document goed Jiri Nossent

Goedkeuring

Afdelingshoofd:	Bellafkih, K.	Getekend door:Abdelkarim Bellafkih (Sign Getekend op:2023-01-30 14:29:08 +01:0 Reden:Ik keur dit document goed Abdelkarim Bellafkih

Abstract

In het kader van de doelstelling uit het departementale ondernemingsplan van 2017, werden de afgelopen jaren de mogelijkheden voor het gebruik van Big Data en Data-mining technieken, en meer bepaald Machine Learning technieken, voor het Waterbouwkundig Laboratorium geëxploreerd. Deze innovatieve technieken vinden immers meer en meer ingang in ons vakgebied en gezien de grote hoeveelheid beschikbare data bij het WL, zou dit dus op termijn een sterke meerwaarde kunnen betekenen.

Voor het exploreren van de mogelijkheden hiervoor en voor de nodige kennisopbouw, werd er enerzijds samengewerkt met enkele universitaire groepen – waarbij studenten aan de slag gingen onder leiding van een expert in het vakgebied – en werd ook een case uitgewerkt in samenwerking met het MOW datalab. De resultaten van deze verschillende projecten worden samengevat in dit rapport. Naast enkele veelbelovende resultaten, werden ook nog een aantal hiaten en moeilijkheden geïdentificeerd. Het is dan ook duidelijk dat dit rapport geen eindpunt kan en mag zijn voor het gebruik van deze technieken, aangezien de mogelijkheden zeer groot zijn en er nog veel progressie mogelijk is. Het werk wordt daarom verder gezet onder Permanente Activiteit 0066.

Inhoudstafel

Lijst van de tabellen

Lijst van de figuren

1 Inleiding

In 2017 stelde het departementale ondernemingsplan voorop dat het Waterbouwkundig Laboratorium in de toekomst de lange termijn hoogwatervoorspellingen zou kunnen valideren en verbeteren door het gebruik van Big Data en Data-mining technieken. Deze innovatieve technieken vinden immers meer en meer ingang in ons vakgebied en gezien de grote hoeveelheid beschikbare data bij het WL, zou dit dus op termijn een sterke meerwaarde kunnen betekenen.

Door de beperkte aanwezige kennis binnen het Waterbouwkundig Laboratorium op het gebied van Big Data, Data-mining en Machine Learning, en gezien de brede waaier aan technieken die hiervoor gebruikt kunnen worden, werd er voor gekozen om een exploratie uit te voeren om mogelijke opportuniteiten te definiëren en vast te leggen door contacten met experts en kennisopbouw.

In eerste instantie werd hierbij gekozen om enkele studiedagen bij te wonen om de nodige kennis op te bouwen. Daarnaast werd met succes een voorstel voor een MSc thesis gelanceerd bij verschillende universitaire vakgroepen die gespecialiseerd zijn in dit onderwerp. De dataset die we hierbij ter beschikking stelden, wekte bij verschillende vakgroepen interesse op, waardoor er nog bijkomende projecten met studenten werden opgestart en er ook interesse bleek om verder onderzoek hierop uit te voeren. Er dient hierbij wel vermeld te worden dat in de nasleep van de corona-pandemie de concrete universitaire contacten verminderden en de verdere uitwerking uitdoofde.

Verder werd er in samenwerking met enkele Machine Learning experten van het MOW datalab een project opgestart om een concrete toepassing voor onze voorspellingen te ontwikkelen. Deze samenwerking resulteerde in een inzetbare tool die nuttig kan zijn voor de HIC-permanentie.

Een overzicht van de activiteiten die plaats vonden in het kader van project 17_070 (Exploreren mogelijkheden Big Data en Data-mining) wordt gegeven in Tabel 1. De verdere ontwikkelingen rond dit thema worden verder gezet in permanente activiteit PA066.

Tabel 1 – Activiteiten binnen project 17_070

Datum	Activiteit	Verwijzing
12/12/2017	Studiedag: Big Data en de overheid (KU Leuven)	-
2018-2019	MSc Thesis Bob De Clercq, UGent	Project 18_101 - §2.1
2018-2019	MSc thesis Jonathan Bokungu, VUB	Project 18_100 - §2.2
09/12/2019	Seminarie EAO - AI/Big Data	-
2020	Vervolgonderzoek Nicolas Dewolf, Ugent	-
2020-2021	Project Waterstanden MOW Datalab	§ 3
2020-2021	LINMA project, UCL	§ 2.3

2 Thesisvoorstel

Om een eerste exploratie van de mogelijkheden van Machine Learning technieken te doen, schreven een groep onderzoekers van het Waterbouwkundig Laboratorium (Paul Vanderkimpen, Maarten Deschamps, Joris Vanlede en Jiri Nossent) een thesisvoorstel in het kader van het maken en verbeteren van de voorspellingen van waterstanden langsheen de Schelde. Deze voorstellen werden ingediend bij gespecialiseerde vakgroepen aan verschillende universiteiten (UA – Prof. Bart Goethals; UGent – Prof. Willem Waegeman; VUB – Prof. Nikos Deligiannis; KUL – Prof. Patrick Willems). Bij de lancering van het onderwerp, werd dit door twee studenten gekozen voor hun afstudeerwerk (zie §2.1 en §2.2).

"Big Data" en "Machine Learning" voor betere tijverwachtingen langs de Schelde

Het Hydrologisch Informatiecentrum (HIC) van het Waterbouwkundig Laboratorium is verantwoordelijk voor het opstellen van verwachtingen voor de hoogwater- en laagwaterstanden langs de Schelde (bv. te Antwerpen). Zeker bij verhoogde waterstanden onder invloed van springtij en/of storm is het belangrijk dat de verwachtingen zo accuraat mogelijk gebeuren. Er kunnen immers op basis van de verwachtingen maatregelen genomen worden om de impact van overstromingen te verkleinen en zo de bevolking te beschermen.

Een heel aantal factoren die de effectieve waterstand langsheen de Schelde beïnvloeden zijn bekend. Zo weten we zeker dat de waterstand op de Noordzee en de wind langsheen de Schelde een bepalende invloed hebben op de uiteindelijke waterstanden. Onze ervaring leert ons echter dat er ook factoren zijn die we nog niet kennen of kunnen inschatten, waardoor de waterstanden soms hoger of lager zijn dan verwacht. Zo is de invloed van de huidige staat en de geschiedenis van het systeem op het volgende hoogwater nog niet voldoende gekend.

Het doel van deze thesis is daarom ***het (1) identificeren en (2) opstellen van relaties tussen het verschil in waterstand te Vlissingen en resp. Terneuzen, Hansweert, Prosperpolder en Antwerpen enerzijds, en alle mogelijke invloedsfactoren anderzijds, op basis van "Big Data" en "Machine Learning" technieken en algoritmes****. Een dergelijke relatie kan immers nieuwe inzichten verschaffen in de factoren die de waterstand beïnvloeden en hulp bieden bij het opstellen van de verwachtingen van het HIC (in het bijzonder bij storm-condities).*

De benodigde data (bv. waterstanden, windsnelheden, windrichtingen,…) worden aangeleverd door het Waterbouwkundig Laboratorium en de uitvoering van het onderzoek gebeurt ook in samenwerking met het Waterbouwkundig Laboratorium.

Big data and machine learning techniques to improve the forecast of water levels

Every day, the Hydrological Information Centre (HIC) at Flanders Hydraulics Research (Flemish Government) has to deliver forecasts for the (high and low) water levels along the tidally influenced River Scheldt (e.g. near Antwerp). In particular for very high-water level conditions (e.g. due to springtide and/or storm events), it is of the utmost importance that the forecasts are reliable and precise, in order to be able to take precautions against the possible impact of floods and, hence, protect the population.

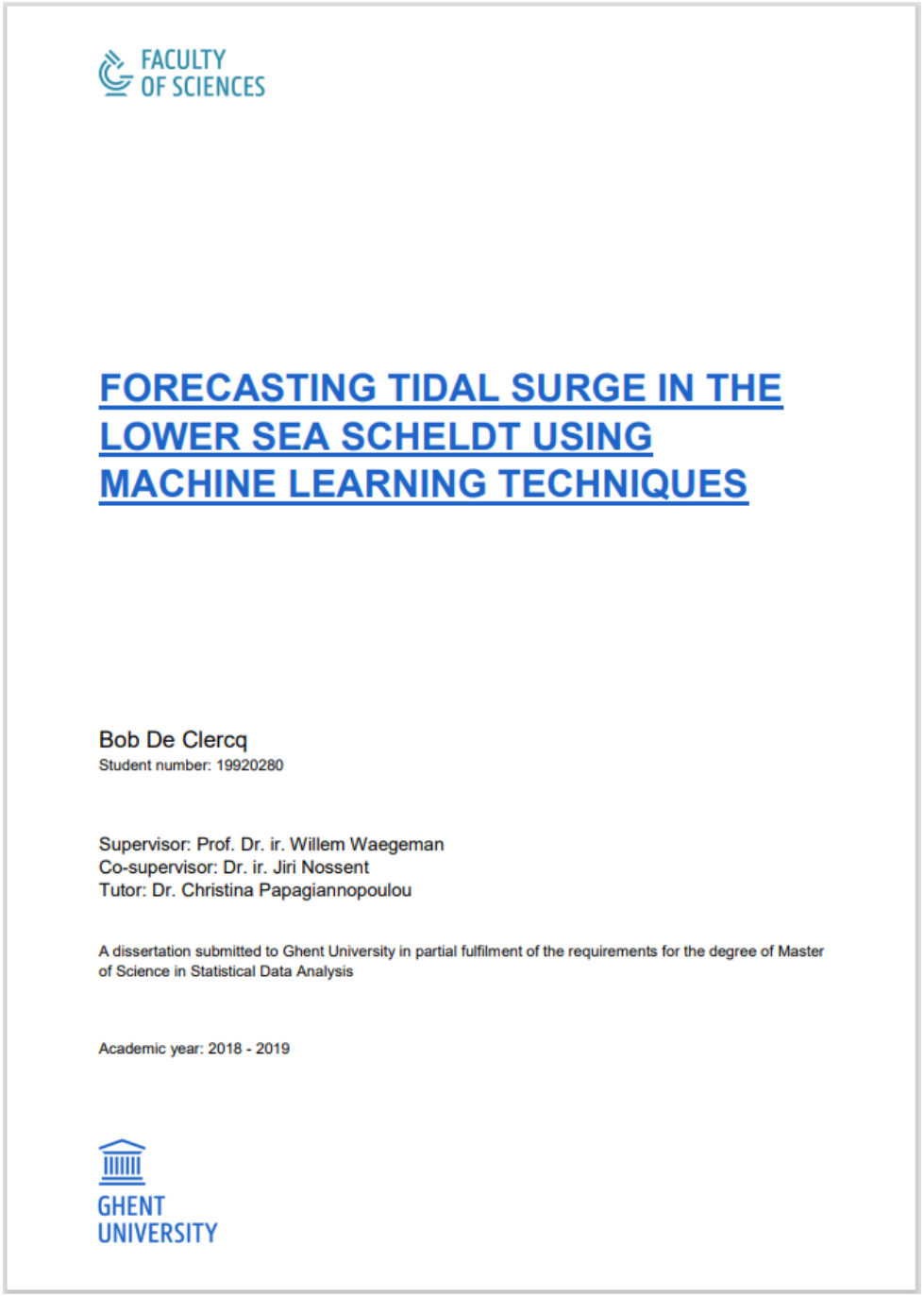
The actual water levels in the tidal area of the Scheldt basin are influenced by a number of factors. E.g. the water levels at the North Sea and the wind along the River Scheldt have an important impact on the final water levels. On the other hand, it is also observed that other, unknown factors have an influence on the system, as the water levels are sometimes over- or underestimated. As an example, the influence of the current state and history of the system on the next high water is poorly understood.

The objective of this MSc thesis is therefore ***to use big data (e.g., proprietary and online data mining) and machine learning (deep learning) techniques and algorithms to (1) identify and (2) define a relation between on the one hand, the difference in water levels near Vlissingen and near resp. Terneuzen, Hansweert, Prosperpolder and Antwerp, and on the other hand, all possible factors that might influence these differences.*** *Such relation can provide new insights in the system and the factors that are determining the water levels in the Scheldt basin and can be an interesting tool for improving the forecasts of the Hydrological Information Centre (in particular during storm-conditions).*

All necessary data (e.g. water levels, wind speed, wind directions,…) will be provided by Flanders Hydraulics Research. Moreover, the research will be carried out in collaboration with Flanders Hydraulics Research.

2.1 MSc thesis Bob De Clercq, UGent – *Forecasting Tidal Surge in the Lower Sea Scheldt using Machine Learning Techniques*

Onder begeleiding van Prof. Willem Waegeman voerde Bob De Clercq tijdens het academiejaar 2018-2019 zijn thesisonderzoek uit met als titel “Forecasting Tidal Surge in the Lower Sea Scheldt using Machine Learning Techniques” (Figuur 1) in het kader van de opleiding "Statistical Data Analysis” binnen de faculteit Wetenschappen van de UGent. Dit onderzoek werd bij het WL geregistreerd als project 18_101, child-project van project 17_070. De uiteindelijke thesis kan onder dit project gevonden worden.



FACULTY OF SCIENCES

FORECASTING TIDAL SURGE IN THE LOWER SEA SCHELDT USING MACHINE LEARNING TECHNIQUES

Bob De Clercq
Student number: 19920280

Supervisor: Prof. Dr. ir. Willem Waegeman
Co-supervisor: Dr. ir. Jiri Nossent
Tutor: Dr. Christina Papagiannopoulou

A dissertation submitted to Ghent University in partial fulfilment of the requirements for the degree of Master of Science in Statistical Data Analysis

Academic year: 2018 - 2019

GHENT UNIVERSITY

Figuur 1 – Titelblad van de MSc thesis van Bob De Clercq

Bij het onderzoek werden verschillende technieken getest, waarbij de focus vooral lag bij het voorspellen van de opzet met een zichttijd van 24u. De techniek die hierbij de beste resultaten gaf werd ook gebruikt om een model met zichttijd 6u op te zetten.

Een eerste reeks modellen die gebruikt werden om voorspellingen te maken met een zichttijd van 24u, vallen onder de noemer “Lineaire modellen”, en meer bepaald “lineaire regressie modellen”:

- Ordinary linear regression (basis lineaire regressie waarbij een verband gezocht wordt tussen een afhankelijke variabele en één of meerdere onafhankelijke variabelen)
- Lasso linear regression (“least absolute shrinkage and selection operator”: specifieke lineaire regressie met regularisatie om een eenvoudigere oplossing te verkrijgen)
- Ridge linear regression (specifieke lineaire regressie voor variabelen met hoge correlatie)
- Elastic-net linear regression (combinatie van Lasso en Ridge lineaire regressie)

De resultaten die verkregen werden voor deze methodes worden weergegeven in Tabel 2.

Tabel 2 – Resultaten van de Lineaire regressie modellen met 24u zichttijd (RMSE in [m]) (De Clercq, 2019)

	training data		test data	
	RMSE	R^2	RMSE	R^2
Ordinary linear regression	0.214	0.450	0.236	0.261
Lasso regression	0.214	0.445	0.235	0.269
Ridge regression	0.214	0.445	0.235	0.269
Elastic-net regression	0.214	0.445	0.235	0.269

Binnen de categorie van niet-lineaire modellen werden nog vier andere technieken toegepast:

- Random Forest Regression (regressie gebaseerd op een combinatie van beslissingsbomen)
- Least-Squares Support Vector Regression (regressive gebaseerd op gesuperviseerde patroonherkenning)
- Extreme Learning Machines (gebaseerd op een niet cyclisch neuraal netwerk)
- Multiple Layer Perceptron (een neuraal netwerk met meerdere lagen)

De resultaten die verkregen werden voor deze methodes worden weergegeven in Tabel 3.

Tabel 3 – Resultaten van de niet-lineaire modellen met 24u zichttijd (RMSE in [m]) (De Clercq, 2019)

	training data		test data	
	RMSE	R^2	RMSE	R^2
Random forests	0.030	0.989	0.180	0.568
LS-SVR	0.177	0.623	0.225	0.326
ELM	0.218	0.428	0.233	0.284
MLP	0.218	0.428	0.235	0.271

Op basis van bovenstaande resultaten is dus duidelijk dat de Random Forest Regressie de beste resultaten oplevert. Aangezien echter vooral de extreem hoge waarden interessant zijn voor het gebruik van een

dergelijk modellen, werd nog een bijkomende “POT weging” toegepast om het model op te bouwen, waarbij de 5% hoogste waterstanden in de dataset (in dit geval boven 5,39 mTAW) een hoger gewicht krijgen. Het resultaat hiervan (met verschillende gewichtsfactoren) is te zien in Tabel 4. Hierbij zijn “RMSE” en “R^2” de criteria voor alle hoogwaters, terwijl “RMSE POT” en “R^2 POT” alleen de hoogste hoogwaters evalueert.

Tabel 4 – Resultaten van de gewogen Random Forest regressie met 24u zichttijd (RMSE in [m]) (De Clercq, 2019)

	training data				test data			
weight	RMSE	RMSE POT	R^2	R^2 POT	RMSE	RMSE POT	R^2	R^2 POT
1	0.030	0.037	0.989	0.986	0.180	0.184	0.568	0.639
30	0.070	0.000	0.940	1.000	0.181	0.171	0.567	0.688
60	0.096	0.000	0.888	1.000	0.180	0.172	0.568	0.684
90	0.113	0.000	0.846	1.000	0.181	0.169	0.567	0.693

Tot slot werd nog een model met een zichttijd van 6 uur getraind op basis van de Random Forest regressie. De resultaten hiervan worden weergegeven in Tabel 5. Zoals te verwachten viel, leveren de voorspellingen voor de test data betere resultaten op met een kortere zichttijd.

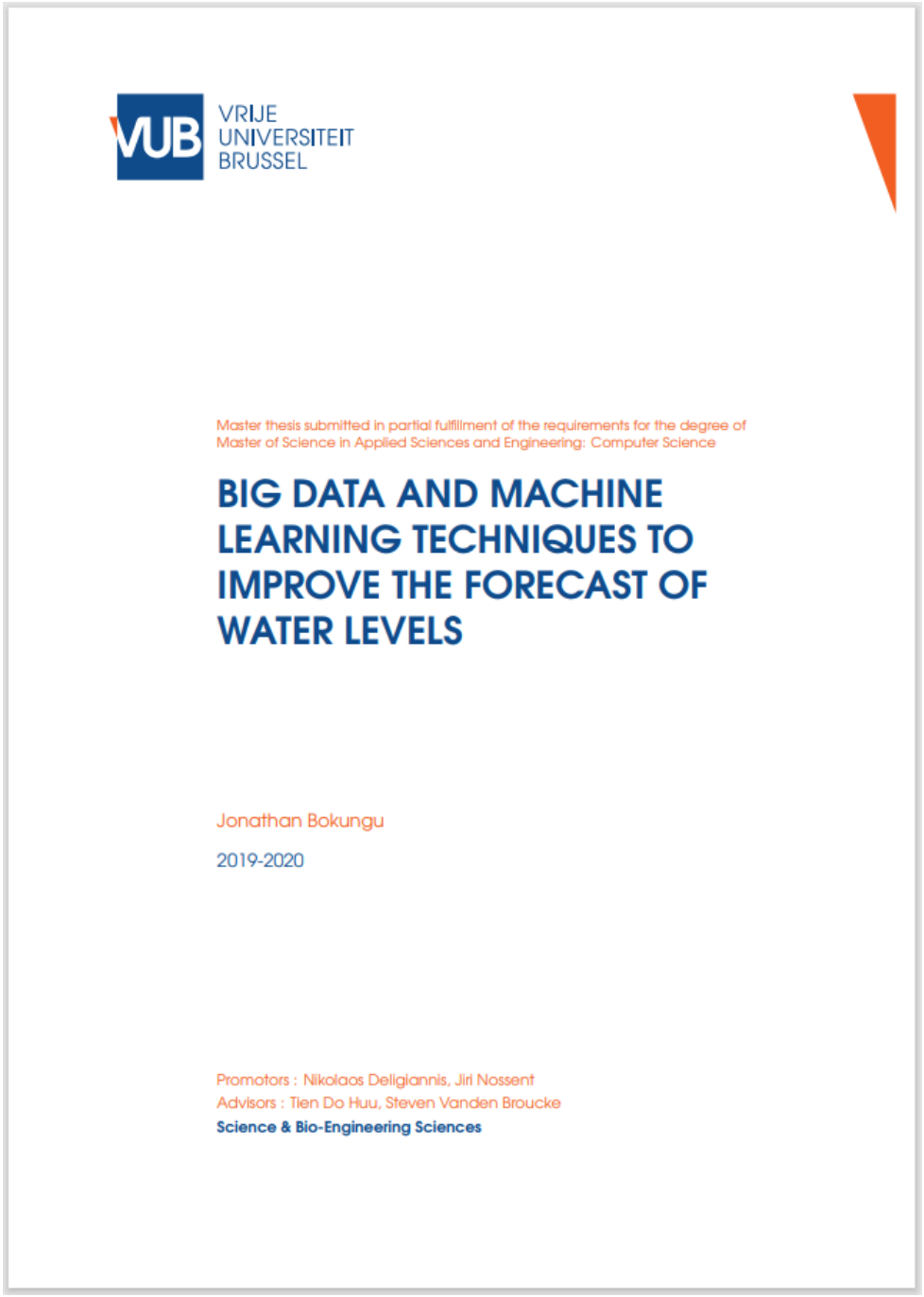
Tabel 5 – Resultaten van de gewogen Random Forest regressie met 6u zichttijd (RMSE in [m]) (De Clercq, 2019)

	training data				test data			
weight	RMSE	RMSE POT	R^2	R^2 POT	RMSE	RMSE POT	R^2	R^2 POT
1	0.036	0.041	0.984	0.983	0.157	0.162	0.670	0.712
30	0.075	0.000	0.933	1.000	0.158	0.152	0.668	0.747

Om de betrouwbaarheid en onzekerheden van dergelijke modellen in te schatten werd in deze thesis ook nog een “conformal prediction framework” gebruikt om de veronderstelling van normaal verdeelde data te vermijden.

2.2 MSc thesis Jonathan Bokungu, VUB – *Big Data and Machine Learning Techniques to improve the Forecast of Water Levels*

Onder begeleiding van Prof. Nikos Deligiannis voerde Jonathan Bokungu tijdens het academiejaar 2018-2019 zijn thesisonderzoek uit met als titel “Big Data and Machine Learning Techniques to improve the Forecast of Water Levels” (Figuur 2) in het kader van de opleiding "Computer Science” binnen de faculteit Ingenieurswetenschappen van de VUB. Dit onderzoek werd bij het WL geregistreerd als project 18_100, child-project van project 17_070. De uiteindelijke thesis kan onder dit project gevonden worden.



Figuur 2 – Titelblad van de MSc thesis van Jonathan Bokungu

Bij dit onderzoek werden verschillende “deep learning” technieken getest en gecombineerd. De meeste hiervan vallen onder de categorie “Recurrent neural networks” (RNN) (neurale netwerken die rekening houden met onderlinge afhankelijkheid van waarden in een reeks):

- Stacked Long Short-Term Memory (LSTM) modellen (RNN modellen die op efficiënte manier oude informatie kunnen gebruiken)
- Gated Recurrent Unit (GRU) modellen (Een vereenvoudigde LSTM met mogelijkheden om ontbrekende data te omzeilen)
- Encoder-decoder modellen (Combinatie van twee RNNs, waarbij de eerste de input processed en de tweede het resultaat van deze processing omzet naar een output)

Een andere categorie van technieken die gebruikt werden zijn statistische modellen en meer bepaald:

- (Seasonal) AutoRegressive Integrated Moving Average ((S)ARIMA) modelllen (modellen gebaseerd op autoregressie, de sterke onderlinge afhankelijkheid van opeenvolgende waarden in een rijdreeks, en de verschuivend venster om het gemiddelde te bepalen)

Een uitgebreide beschrijving van deze technieken kan geraadpleegd worden in (Bokungu, 2020).

De resultaten die verkregen werden voor de verschillende gecombineerde methodes voor het jaar 2014 worden weergegeven in Tabel 6 en Tabel 7 en geven het verschil (in [m]) tussen de door het ML model gemodelleerde waarde en de effectieve waarde.

Tabel 6 – Mean Absolute Error ([m]) voor de verschillende modellen (Bokungu, 2020)

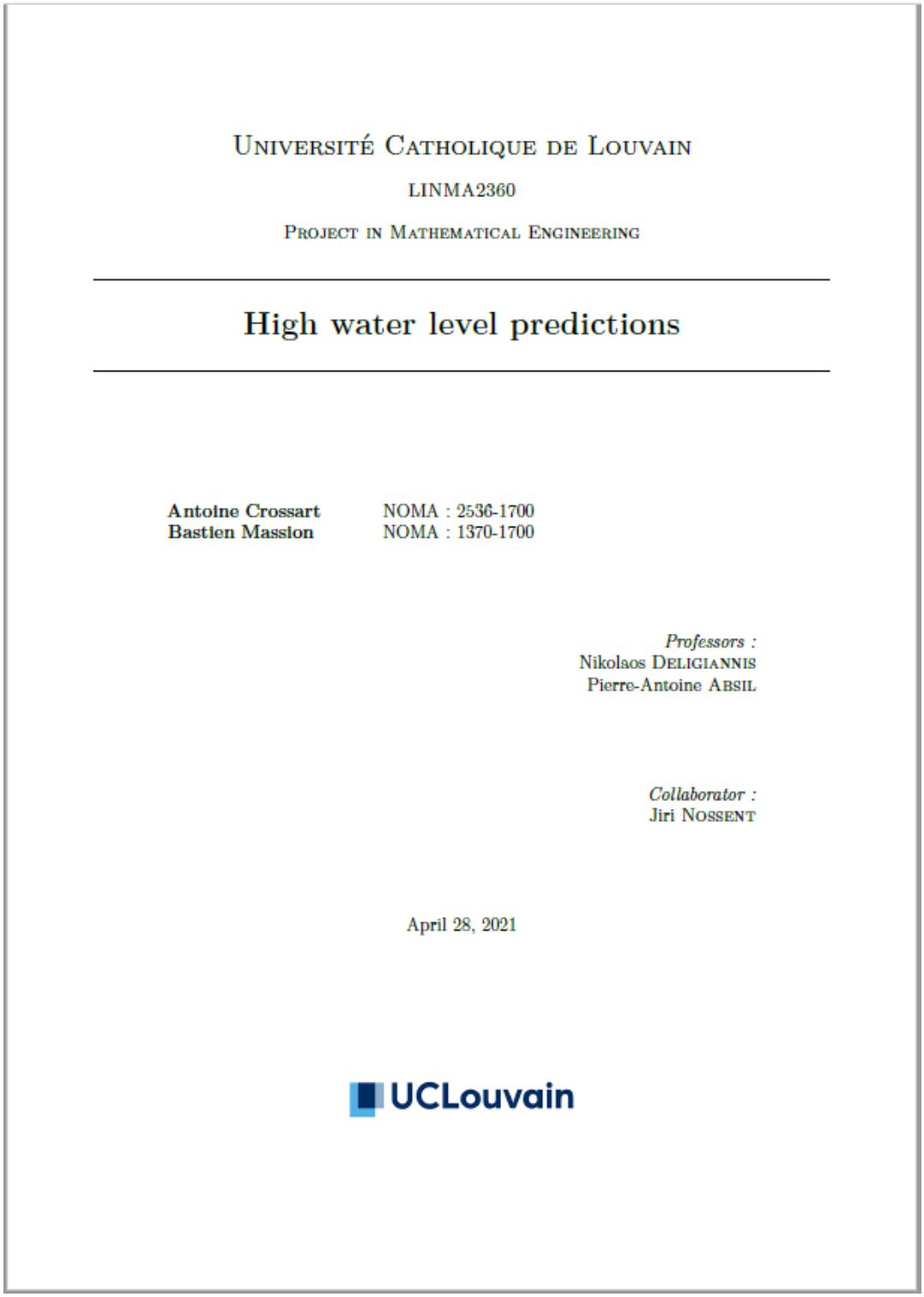
Models	Antwerp	Vlissingen	Astronomical
Baseline LSTM	0.166	0.118	0.111
Baseline GRU	0.164	0.120	0.122
Encoder-Decoder LSTM	0.131	0.118	0.148
Encoder-Decoder GRU	0.168	0.150	0.159
Encoder-Decoder 1D-CNN - LSTM	0.126	**0.103**	0.123
Encoder-Decoder Conv2D - LSTM	**0.108**	0.104	**0.087**
ARIMA	*0.110*	**0.103**	*0.094*

Tabel 7 – Root Mean Squared Error ([m]) voor de verschillende modellen (Bokungu, 2020)

Models	Antwerp	Vlissingen	Astronomical
Baseline LSTM	0.199	**0.148**	0.131
Baseline GRU	0.232	0.160	0.164
Encoder-Decoder LSTM	**0.178**	0.163	0.166
Encoder-Decoder GRU	0.202	0.164	0.198
Encoder-Decoder 1D-CNN - LSTM	0.194	0.169	0.143
Encoder-Decoder Conv2D - LSTM	**0.178**	0.163	**0.135**
ARIMA	*0.180*	*0.168*	*0.140*

2.3 Project LINMA, UCL (Antoine Crossart & Bastien Massion) – *High water level predictions*

Onder begeleiding van Prof. Pierre-Antoine Absil and Prof. Nikos Deligiannis voerden Antoine Crossart en Bastien Massion een klein project uit voor de cursus “Mathematical Engineering” aan de UC Louvain-La-Neuve. Het rapport van dit project kan gevonden worden op de projectsite van project 17_070.



UNIVERSITÉ CATHOLIQUE DE LOUVAIN

LINMA2360

PROJECT IN MATHEMATICAL ENGINEERING

High water level predictions

Antoine Crossart NOMA : 2536-1700
Bastien Massion NOMA : 1370-1700

Professors :
Nikolaos DELIGIANNIS
Pierre-Antoine ABSIL

Collaborator :
Jiri NOSSENT

April 28, 2021

UCLouvain

Figuur 3 – Titelblad van het projectrapport van Antoine Crossart en Bastien Massion

In het kader van dit project werden twee Machine Learning technieken toegepast om een model met een zichttijd van 12u op te stellen:

- Vector Autoregressive Model (VAR) (model gebaseerd op autoregressie)
- Long Short-Term Memory (LSTM) model (RNN dat op efficiënte manier oude informatie kan gebruiken)

Er werd ook getracht om een link te maken tussen de waterstand in Antwerpen en de waterstanden opwaarts en afwaarts van Antwerpen.

De resultaten van het Vector Autoregressive Model worden weergegeven in Tabel 8. Bij de opbouw van het model werden aanvankelijk alleen de gegevens voor Antwerpen gebruikt. In een later stadium werden ook de metingen voor respectievelijk Vlissingen ("Vliss"), Hemiksem ("Hem"), Temse ("Temse"), Zandvliet ("Zand"), alsook alle reeksen met waterstanden ("All cities") en de windgegevens voor Hansweert ("wind") toegevoegd. Het "True percentage" is hierbij het percentage juist voorspelde hoogwaters (i.e. in deze case waterstanden boven 6 mTAW die voorspeld werden die ook effectief optraden) t.o.v. het aantal hoogwaters of tijcycli. Het "False percentage" is het percentage "false positives" (i.e. waterstanden boven 6 mTAW die voorspeld werden terwijl deze niet optraden) t.o.v. het aantal hoogwaters of tijcycli. Het toevoegen van de wind zorgt voor een duidelijke verbetering van het aantal correct voorspelde hoogwaters.

Tabel 8 – Resultaten voor het Vector Autoregressive Model (Crossart & Massion, 2021)

City	Antwerp	A+Vliss	A+Hem	A+Temse	A+Zand	All cities	A+wind
MSE	0.0093	0.0111	0.0093	0.0093	0.0117	0.0146	0.020
True percentage	41.35	43.61	41.35	41.35	46.61	41.35	54.82
False percentage	3.01	3.01	3.01	3.01	2.25	1.50	3.61

De resultaten van het Long Short-Term Memory model (RNN) worden weergegeven in Tabel 9. Deze zijn duidelijk minder sterk dan deze voor VAR.

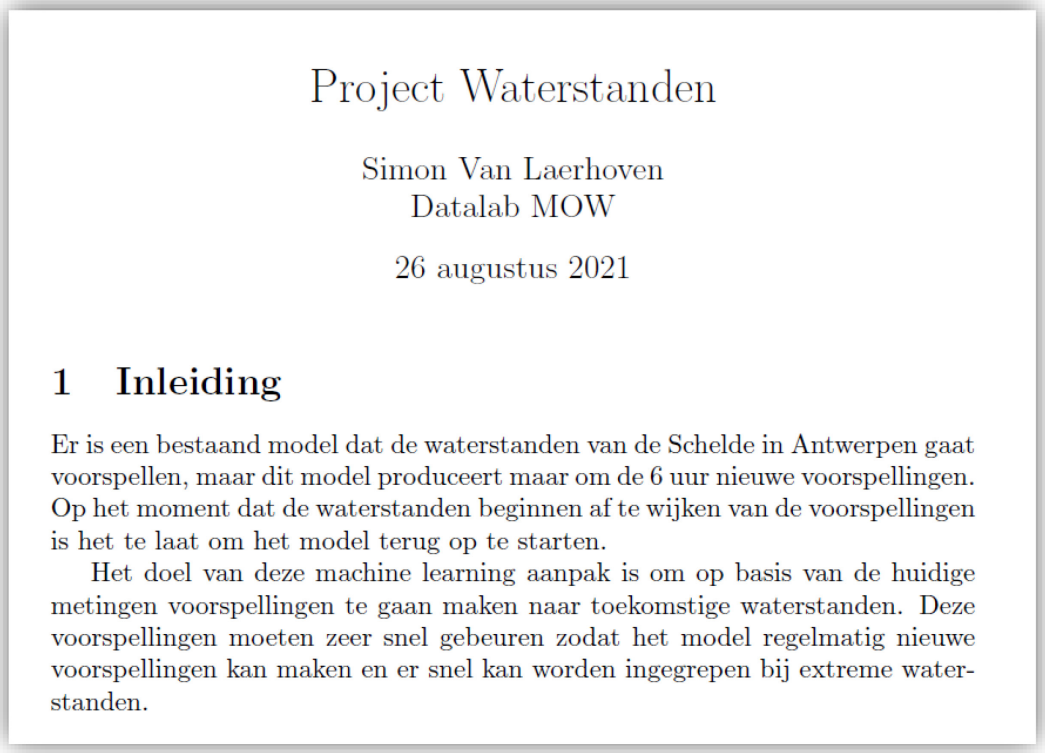
Tabel 9 – Resultaten voor het Long Short-Term Memory model (Crossart & Massion, 2021)

City	Antwerp	A+Vliss	A+Hemiksem	A+Temse	A+Zandvliet	All city	A+wind
MSE	0.054	0.055	0.056	0.062	0.053	0.053	0.050
True percentage	32.69	34.95	30.42	33.33	29.45	28.80	35.54
False percentage	15.86	17.48	16.18	17.48	13.59	13.27	15.66

3 Samenwerking MOW Datalab

Na het volgen van het seminarie “EAO - AI/Big Data”, werd er een samenwerking opgezet met het MOW Datalab om een concrete case rond het voorspellen van waterstanden op de Schelde op te zetten. Hoewel er een heel aantal concrete, uiteenlopende ideeën waren (bv. het onderzoeken van verbanden tussen de oriëntatie van bepaalde “reaches” van de Zeeschelde en de evolutie van de waterstanden langsheen de rivier), werd er gekozen om een specifieke tool te ontwikkelen, die enerzijds een concrete toegevoegde waarde kon bieden aan het werk van de HIC-permanentie en anderzijds haalbaar zou zijn binnen de beperkingen die eerder al werden vastgesteld bij het werken met Machine Learning technieken (e.g. beperkte tijdshorizon, beperkte complementariteit tussen gemeten en voorspelde input,…).

Concreet werd er gewerkt aan de ontwikkeling van een tool om, met een zeer korte rekentijd, een korte termijn voorspelling te genereren voor de waterstand in Antwerpen. In het verleden gebeurde het immers al dat gedurende een storm de gemeten waterstanden in Oostende en/of Vlissingen een heel stuk hoger uitkwamen dan de waarde die vooraf was voorspeld. Aangezien de verwachte maximale waterstand voor Antwerpen grotendeels gebaseerd is op de voorspelde waterstanden voor Oostende en Vlissingen binnen dezelfde getijgolf, leidt een dergelijke onderschatting van deze laatste twee ook onvermijdelijk tot een onderschatting van de maximale waterstand in Antwerpen. Door op basis van de gemeten waterstanden voor Oostende en Vlissingen nog een snelle voorspelling te kunnen maken voor Antwerpen met een Machine Learning model (het maximale hoogwater in Antwerpen treedt ongeveer 2u na het maximale hoogwater in Vlissingen op), zou het dan mogelijk zijn om potentiële problemen bij een verhoogde waterstand alsnog te identificeren. Een dergelijke berekening is quasi onmogelijk met onze traditionele hydrodynamische modellen, aangezien deze een grotere rekentijd vergen.



Project Waterstanden

Simon Van Laerhoven
Datalab MOW

26 augustus 2021

1 Inleiding

Er is een bestaand model dat de waterstanden van de Schelde in Antwerpen gaat voorspellen, maar dit model produceert maar om de 6 uur nieuwe voorspellingen. Op het moment dat de waterstanden beginnen af te wijken van de voorspellingen is het te laat om het model terug op te starten.

Het doel van deze machine learning aanpak is om op basis van de huidige metingen voorspellingen te gaan maken naar toekomstige waterstanden. Deze voorspellingen moeten zeer snel gebeuren zodat het model regelmatig nieuwe voorspellingen kan maken en er snel kan worden ingegrepen bij extreme waterstanden.

Figuur 4 – Rapport van het “Project Waterstanden” van het MOW datalab

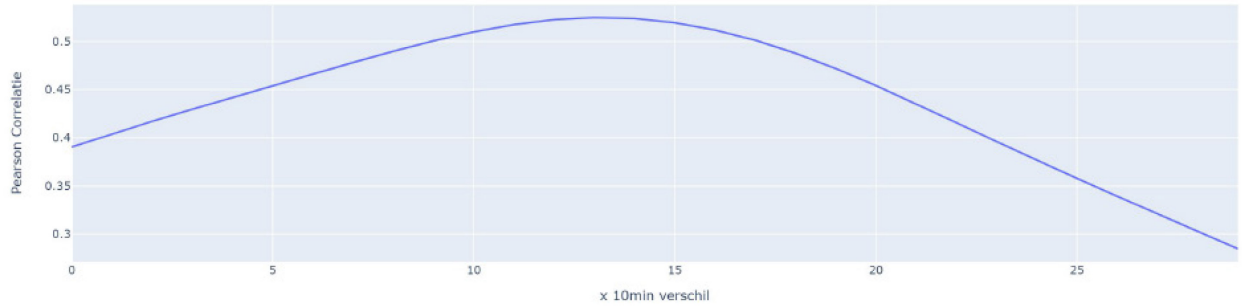
3.1 Data

Bij het opbouwen van het Machine Learning model door het MOW Datalab, werden dezelfde input data gebruikt als bij het werk dat door de academische groepen werd verricht (zie bijlage 1). Het gaat hier meer bepaald om waterstanden (metingen, astronomische voorspellingen en gewone voorspellingen), debieten, wind snelheden, wind richtingen, bijkomende meteorologische parameters,… Deze data worden voor deze toepassing echter in real-time ingeladen in het systeem om de korte termijn voorspelling mogelijk te maken.

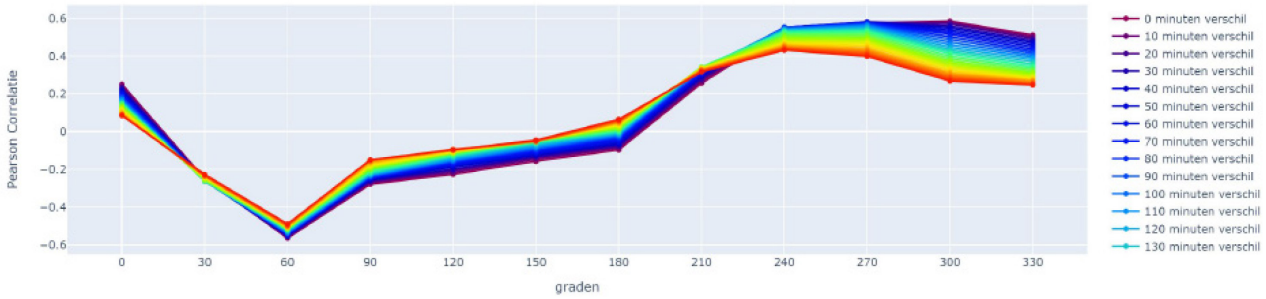
Ter voorbereiding van de berekeningen worden de tijdreeksen geresampled naar een tijdstap van 10 minuten en worden bovendien ontbrekende gegevens aangevuld door lineaire interpolatie.

3.2 Correlaties

Om een inschatting te maken van de factoren die de grootste invloed hebben op de waterstanden langs de Schelde in Antwerpen, werd er eerst onderzocht welke correlaties er bestaan tussen de verschillende variabelen. Zo is het duidelijk dat de correlatie tussen de opzet in Vlissingen en de opzet in Antwerpen op een tijdsverschil van ongeveer 2u ligt (Figuur 5) en toont Figuur 6 dat zowel de grootste positieve correlatie als de meest negatieve correlatie tussen de windsnelheid in Vlissingen en de opzet in Antwerpen gevonden wordt bij 0 minuten tijdsverschil tussen beide. In Figuur 6 is bovendien duidelijk zichtbaar dat er voor windrichtingen 30-90° (NNO-O) een negatieve correlatie bestaat en voor 210-360° (ZW-N) een positieve.



Figuur 5 – Correlatie tussen de opzet in Vlissingen en Antwerpen voor verschillende tijdsverschuivingen (Van Laerhoven, 2021)



Figuur 6 – Correlatie tussen de windsnelheid in Vlissingen en de opzet te Antwerpen per snede van 30° voor de windrichtingen en voor verscheidene tijdsverschillen (Van Laerhoven, 2021)

Tabel 10 geeft bovendien ook nog een overzicht van de sterkste correlaties die gevonden werden bij de analyse van de verschillende variabelen t.o.v. de opzet te Antwerpen (i.e. de doelvariabele). Een zelfde analyse voor waterstanden van meer dan 6,0 mTAW wordt weergegeven in Tabel 11.

Tabel 10 – Overzicht van de sterkst waargenomen correlaties met de opzet in Antwerpen (=doelvariabele) (Van Laerhoven, 2021)

Variabele	Correlatie	Correlatietype	Transformatie
oostende surge deviation	0.7658	pearson	3 uur voor doelvariabele
vlissingen - wind speed	0.5865	pearson	op -60° zelfde tijd als doelvariabele
vlissingen - wind speed	0.5823	pearson	op -90° 30 min voor doelvariabele
vlissingen - wind speed	-0.5638	pearson	op +60° zelfde tijd als doelvariabele
vlissingen - wind speed	0.5533	pearson	op -120° 60 min voor doelvariabele
vlissingen surge deviation	0.5247	pearson	2u10min voor doelvariabele
Zeebrugge - Air Pressure	-0.3138	spearman	13u50min voor doelvariabele
Eppegem Zenne - River Discharge	0.2616	pearson	50 min voor doelvariabele
Grobbendonk Troon Kleine Nete - River Discharge	0.2191	pearson	zelfde tijd als doelvariabele
Dendermonde Dender - River Discharge	0.2060	pearson	2u voor doelvariabele

Tabel 11 – Overzicht van de sterkst waargenomen correlaties bij een waterstand > 6,0 mTAW Antwerpen (Van Laerhoven, 2021)

Variabele	Correlatie	Correlatietype	Transformatie
oostende surge deviation	0.8416	pearson	3 uur voor doelvariabele
vlissingen surge deviation	0.8445	pearson	1u50min voor doelvariabele
Eppegem Zenne - River Discharge	0.3186	spearman	50 min voor doelvariabele
Zeebrugge - Air Pressure	-0.2568	spearman	9u40min voor doelvariabele
Grobbendonk Troon Kleine Nete - River Discharge	0.3312	spearman	zelfde tijd als doelvariabele

3.3 Recurrent Neural Network (RNN) model - Tool

Net als bij de projecten die werden uitgevoerd onder leiding van prof. Deligiannis (§2.2, §2.3), werkte het MOW datalab aan een model op basis van een Recurrent Neural Network (RNN). Hierbij werd zowel een voorspellingstijd van 4u als 8u gebruikt. Het trainen van beide modellen duurde ongeveer even lang als deze voorspellingstijden. Het verkregen model geeft als statistische performantie een RMSE (Root Mean Squared Error) van 6.83 cm en een MAE (Mean Absolute Error) van 4.49 cm. Een inschatting van de invloed van verschillende input factoren op de performantie wordt weergegeven in Tabel 12. Wanneer een van deze input factoren uit het model weg gelaten wordt, nemen de RMSE en MAE van het model met de aangegeven waarde toe. Tabel 13 geeft daarnaast een overzicht van de door het model gesimuleerde waarden bij verschillende lead-times voor een aantal historische stormtijen. Het model gebruikt daarvoor de gemeten waterstanden tot het moment dat de berekening gemaakt wordt.

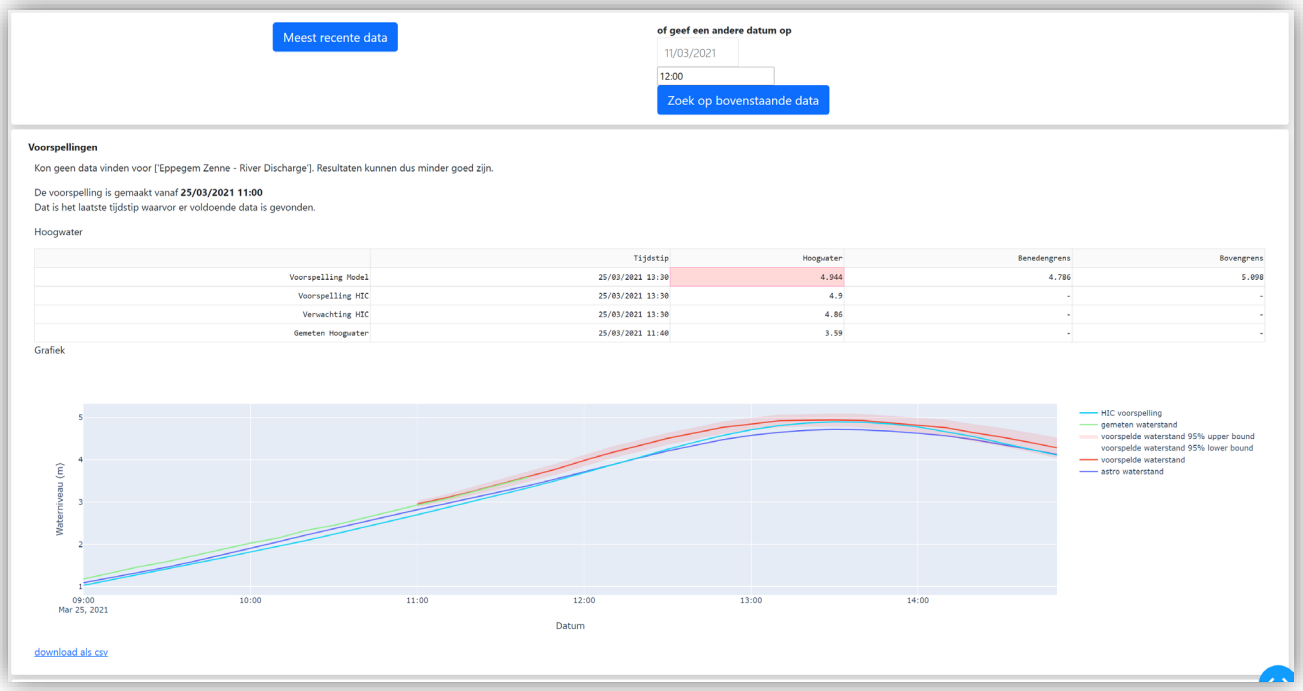
Tabel 12 – Invloed van de verschillende input factoren op de performantie van het model (Van Laerhoven, 2021)

Variabele	$\Delta RMSE(cm)$	$\Delta MAE(cm)$
oostende surge deviation	37.3464	27.8494
vlissingen surge deviation	7.5630	5.7163
Dendermonde Dender - River Discharge	0.0674	0.0608
Eppegem Zenne - River Discharge	0.0070	0.0040
Grobbendonk Troon Kleine Nete - River Discharge	0.0148	0.0064
Hulshout Grote Nete - River Discharge	0.0101	0.0126
Oosterweel-Boven SF Zeeschelde - Water Temperature	0.0570	0.1590
Prosperpolder SF Zeeschelde - Water Temperature	0.1024	0.1397
Zeebrugge - Air Pressure	0.8596	0.0666
Zeebrugge - Air Temperature	0.0187	0.0266
vlissingen - wind x	0.1060	0.0676
vlissingen - wind y	0.9470	0.5329
vlissengen - wind (both)	1.0869	0.6196

Tabel 13 – Modelvoorspellingen voor stormtijen tussen 2009 en 2020

Date	Measured Waterlevel	1h forecast	2h forecast	3h forecast	4h forecast
2009-02-10	6.89	6.946132	7.055139	7.101784	7.00162
2010-02-28	6.67	6.4818	6.689235	6.503091	6.345003
2013-12-06	7.29	7.324412	7.385258	7.470444	7.248328
2013-12-06	6.78	6.758376	6.862355	6.948995	6.724845
2014-10-22	6.78	6.788467	6.789839	6.917774	6.819346
2015-11-28	6.76	6.658748	6.460054	6.335141	5.965687
2017-01-13	6.72	6.757086	6.680636	6.642996	6.032893
2017-01-14	6.65	6.541602	6.642431	6.689338	6.647863
2018-01-03	7.14	7.206635	7.180076	7.198535	7.082269
2020-02-10	6.91	6.996728	7.167834	7.134938	6.933297
2020-02-11	6.72	6.783513	6.806702	6.852696	6.187729
2020-02-11	6.86	6.895877	6.946667	6.858789	6.768521
2020-03-12	6.77	6.764286	6.717209	6.828767	6.817716

Tot slot werd er een tool gebouwd op basis van het ontwikkelde model, die in staat is om real-time voorspellingen te maken op basis van de gemeten waterstanden (Figuur 7). De code van deze tool kan in de toekomst op het netwerk van het Waterbouwkundig Laboratorium geïntegreerd worden.



Figuur 7 – Het dashboard van de tool voor het maken van korte termijn voorspellingen met Machine Learning

4 Conclusies

In het kader van de doelstelling uit het departementale ondernemingsplan van 2017, werden de afgelopen jaren de mogelijkheden voor het gebruik van Big Data en Data-mining technieken, en meer bepaald Machine Learning technieken, voor het Waterbouwkundig Laboratorium geëxploreerd.

De verschillende projecten die uitgevoerd werden onder begeleiding van enkele universitaire groepen, leverden een breed overzicht op van beschikbare Machine Learning technieken en hun mogelijkheden voor het maken van voorspellingen van waterstanden. Er is geen eenduidigheid over welk type van model het beste scoort, al geven voor bepaalde toepassingen Random Forest modellen en modellen gebaseerd op Autoregressie wel de meest veelbelovende resultaten. De afhankelijkheid van de modelopbouw (i.c. de modelleur en de gebruikte datasets) is wel een belangrijke factor om rekening mee te houden.

Daarnaast leverde de samenwerking met het MOW datalab een concrete tool op die inzetbaar is voor het werk van de HIC permanentie. Hieruit bleek dat ook modellen gebaseerd op Recurrente Neurale Netwerken nuttig kunnen zijn.

Hoewel er een aantal beperkingen van Machine Learning technieken werden geïdentificeerd voor onze case (bv. de beperkte performantie voor extreme events door het ontbreken van voldoende datapunten in deze range en de beperkte tijdshorizon die de voorspellingen betrouwbaar houdt), lijken er toch aanzienlijke mogelijkheden te liggen in dit vakgebied. Er kunnen bovendien nog bijkomende, eventueel gerelateerde, cases gedefinieerd worden die met Machine Learning technieken kunnen aangepakt worden. Zo lijken er wel mogelijkheden te liggen in het beter bepalen van de afhankelijkheden van de optredende waterstanden voor verschillende variabelen en kan de tijdshorizon voor de voorspellingen misschien uitgebreid worden door ook de meteorologische voorspellingen op te nemen in het gebouwde voorspellingsmodel. Ook een opsplitsing van het de waterloop in “reaches” met een bepaalde richting en het bepalen van de windinvloed op elk van deze takken lijkt een interessante piste om in de toekomst te bewandelen. Het blijft dan ook aangewezen om in de toekomst verder in te zetten op het ontwikkelen van deze kennis en de bijhorende tools. De verdere ontwikkelingen rond dit thema worden hierbij verder gezet in permanente activiteit PA066.

5 Referenties

Bokungu, J. (2020) Big Data and Machine Learning techniques to improve the forecast of waterlevels. MSc thesis Applied Sciences and Engineering: Computer Science, Vrije Universiteit Brussel

Crossart, A., Massion, B. (2021) High water level predictions, Project in Mathematical Engineering, Université Catholique de Louvain

De Clercq, B. (2019) Forecasting tidal surge in the Lower Sea Scheldt using Machine Learning Techniques. MSc thesis Statistical Data Analysis, Universiteit Gent

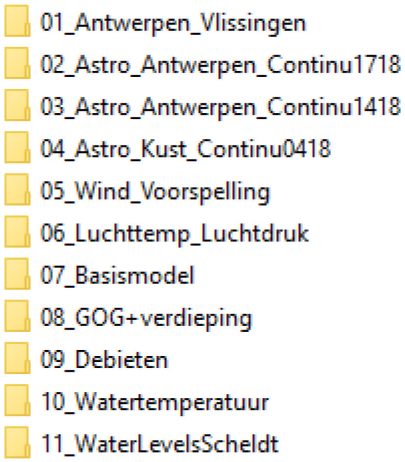
Van Laerhoven, S. (2021) Project Waterstanden. Eindrapport case study MOW datalab.

Bijlage 1: Aangeleverde data

De bulk aan aangeleverde data was voor de verschillende projecten dezelfde. Een algemeen overzicht wordt hieronder gegeven:

1. Waterstandsgegevens voor Antwerpen, Terneuzen, Vlissingen, Bath, Zandvliet, Lillo, Hemiksem, Temse, Dendermonde en Melle
2. Hoog- en Laagwaters voor Antwerpen, Terneuzen, Vlissingen, Bath, Zandvliet, Lillo, Hemiksem, Temse, Dendermonde en Melle
3. Astronomische waterstandsvoorspellingen voor Antwerpen, Vlissingen, Nieuwpoort, Oostende en Zeebrugge
4. Windsnelheid en windrichting data voor Hansweert en Vlakte van de Raan
5. Windsnelheid en windrichting voorspellingen voor Hansweert en Vlakte van de Raan
6. Debieten aan de opwaartse randen te Aarschot, Dendermonde, Eppegem, Melle en Grobbendonk
7. Luchttemperatuur en luchtdruk te Melsele
8. Watertemperatuur te Zandvliet, Lillo en Oosterweel

De data kunnen teruggevonden worden onder projecten 17_070, 18_100 en 18_101. De structuur ziet er als volgt uit:



DEPARTEMENT **MOBILITEIT & OPENBARE WERKEN**
Waterbouwkundig Laboratorium

Berchemlei 115, 2140 Antwerpen
T +32 (0)3 224 60 35
F +32 (0)3 224 60 36
waterbouwkundiglabo@vlaanderen.be
www.waterbouwkundiglaboratorium.be