https://doi.org/10.5194/npg-2021-36
Preprint. Discussion started: 4 January 2022

Nonlinear Processes in Geophysics Discussions

Integrated hydrodynamic and machine learning models for compound flooding prediction in a data-scarce estuarine delta

Joko Sampurno[1,2], Valentin Vallaeys[1], Randy Ardianto[3], Emmanuel Hanert[1,4]

[1]Earth and Life Institute (ELI), Université Catholique de Louvain (UCLouvain), Louvain-la-Neuve, 1348, Belgium
[2]Department of Physics, Fakultas MIPA, Universitas Tanjungpura, Pontianak, 78124, Indonesia
[3]Pontianak Maritime Meteorological Station, Pontianak, 78111, Indonesia
[4]Institute of Mechanics, Materials and Civil Engineering (IMMC), Université Catholique de Louvain (UCLouvain), Louvain-la-Neuve, 1348, Belgium

Correspondence to: Joko Sampurno (joko.sampurno@uclouvain.be, jokosampurno@physics.untan.ac.id)

Abstract. Flood forecasting based on water level modeling is an essential non-structural measure against compound flooding over the globe. With its vulnerability increased under climate change, every coastal area became urgently needs a water level model for better flood risk management. Unfortunately, for local water management agencies in developing countries building such a model is challenging due to the limited computational resources and the scarcity of observational data. Here, we attempt to solve the issue by proposing an integrated hydrodynamic and machine learning approach to predict compound flooding in those areas. As a case study, this integrated approach is implemented in Pontianak, the densest coastal urban area over the Kapuas River delta, Indonesia. Firstly, we built a hydrodynamic model to simulate several compound flooding scenarios, and the outputs are then used to train the machine learning model. To obtain a robust machine learning model, we consider three machine learning algorithms, i.e., Random Forest, Multi Linear Regression, and Support Vector Machine. The results show that this integrated scheme is successfully working. The Random Forest performs as the most accurate algorithm to predict flooding hazards in the study area, with RMSE = 0.11 m compared to SVM (RMSE = 0.18 m) and MLR (RMSE = 0.19 m). The machine-learning model with the RF algorithm can predict ten out of seventeen compound flooding events during the testing phase. Therefore, the random forest is proposed as the most appropriate algorithm to build a reliable ML model capable of assessing the compound flood hazards in the area of interest.

1 Introduction

Compound flooding in low-gradient coastal area is a recognized hazard that can be exacerbated by global warming (Hao and Singh, 2020). Compound flooding hazard is derived by the interaction of storm surge penetration, riverine flooding, and intense rainfall over the area (as the impact of extreme meteorological events) that coincide or nearly coincide (Bilskie and Hagen, 2018). This natural hazard can endanger the population and the coastal area's infrastructures, which have been growing fast in the last decade. Without an appropriate mitigation, the consequences of the hazard can be severe for the coastal environment (Costabile et al., 2013) and the coastal communities both economically (Karamouz et al., 2014) and socially (Comer et al.,

https://doi.org/10.5194/npg-2021-36
Preprint. Discussion started: 4 January 2022


Nonlinear Processes in Geophysics Discussions

2017).

There are various mechanisms driving compound flooding in low-lying urban coastal regions (Santiago-Collazo et al., 2019). First, the water goes up with the tide and sea-level rise due to climate change. Besides, a storm surge also goes on top of it. That water can get into the dry land by wave overtopping. Second, extreme precipitation and high upstream flow discharge can also elevate water in the estuary. Then, the water can overflow and cause flooding as well. These flood pathways are often naturally correlated, driving those mechanisms to occur coincidentally (or in close succession), and hence creating a compound event and worsening the hazard.

Flood forecasting based on water-level modeling in a tidal river area is an essential non-structural measure against compound flooding (Chan, 2015; Tucci and Villanueva, 1999; Mosavi et al., 2018). Water-level modeling could be conducted using a process-based or data-based approach. The process-based approach is more commonly used to tackle the issue (Costabile and Macchione, 2015). However, it requires many assumptions to reduce the complexity and make it computationally tractable. The data-based approach, i.e., machine learning models, has also accurately predicted water level changes without requiring the underlying physical attributes and high computational resources (Choi et al., 2020; Wang and Wang, 2020; Assem et al., 2017). Machine learning is concerned with developing a model that can improve their performance in carrying out a given task over time by learning from examples, with minimal human efforts to instruct them how to do so. Machine learning can enable us to do hypothesis testing and generate confidence bonds in our mitigations. Machine learning models could capture and represent a complex input and output relationship using only historical data (Chen and Asch, 2017).

However, building a flood forecasting model in developing countries is a challenging. Implementing a process-based approach generally requires expensive computational resources (Nayak et al., 2005). Therefore, it is difficult to use for operational management due to the computational resources owned by local agencies is limited. Meanwhile, building a robust machine learning model is required a sufficient amount of data to train the model (Naqa et al., 2018). Unfortunately, the availability of observational data in these areas is also scarce. In addition, applying remote sensing techniques as the solution has also been proposed by some studies (Mokkenstorm et al., 2021; Kabenge et al., 2017; Haq et al., 2012). Nevertheless, due to the limitation of its time resolution, the technique cannot always detect compound flooding, particularly if it occurs in a short time. Besides, remote sensing technique generally is only for detection, monitoring, and validation purposes.

This study attempts to tackle these issues by combining the process-based and data-based approaches. Firstly, we build a hydrodynamic model to run some flood scenarios in a data-scarce estuary. Then, we create machine learning models, which are trained by using the outputs of the hydrodynamic model, to predict the water level and forecast future floods. To obtain a robust machine learning model, we evaluate some machine learning algorithms and select the most accurate one for our application. As a study case, the integrated scheme is implemented in the Kapuas River delta, particularly in the city of Pontianak, to assist the local water management, whose computational resources are limited, and help them assess their compound flood hazard, and mitigate that risk.

https://doi.org/10.5194/npg-2021-36
Preprint. Discussion started: 4 January 2022


Nonlinear Processes in Geophysics Discussions

2 Material and method

2.1 Study area

The Kapuas River is the longest island's river in Indonesia (Goltenboth et al., 2006). Its basin is located in the western part of Borneo Island (Fig. 1). The delta and the estuary of the river are still mostly natural. There are no dams, dykes, or groins on its stream. The river is vital for the local communities as a source of freshwater and transportation. The population along the Kapuas River banks is sparse, except in some parts of the estuary. The most populated area is Pontianak, located along the main branch of the Kapuas River, namely the Kapuas Kecil River. Therefore, the hydrodynamics of the river significantly influences the occurrence of floods in the city. During the last decades, the expansion of palm oil cultivation and forest fires happened massively in the Kapuas water catchment (Semedi, 2014; Jadmiko et al., 2017). These phenomena led to a change in the Kapuas hydrological regime and triggered more intense flooding in the river's floodplains. Combined with global sea level rise these phenomena could lead to more intense and severe flood events in the future.

2.2 Hydrodynamic model description and model setup

We use the multi-scale hydrodynamic model SLIM 2D to simulate compound flooding for different forcing scenarios. Our model can simulate hydro processes along the land-sea continuum, from the river to the ocean (Vallaeys et al., 2018, 2021; Frys et al., 2020; Le et al., 2020b). The model solves the 2D Shallow Water Equations (SWE):

$$\frac{\partial H}{\partial t} + \nabla \cdot \mathrm{U} = \mathrm{R}, \tag{1}$$

$$\frac{\partial \mathrm{U}}{\partial t} + \nabla \cdot \left(\frac{\mathrm{UU}}{H}\right) + f\boldsymbol{e_z} \times \mathrm{U} - \nabla \cdot (\upsilon \nabla \mathrm{U}) = \alpha g H \nabla (H - h) - \frac{Cd}{H^2}|\mathrm{U}|\mathrm{U} + \frac{1}{\rho}\tau_{wind} - \frac{H}{\rho}\nabla P_{atm}\,, \tag{2}$$

where H is the water column height, ∇ is the horizontal gradient operator, $\mathrm{U} = H\bar{\mathbf{u}}$ is the horizontal transport, R is the rainfall, t is the time, $\bar{\mathbf{u}} = (u, v)$ is the depth-averaged horizontal velocity, f is the Coriolis parameter, $\boldsymbol{e_z}$ is the vertical unit vector pointing upward, υ is the horizontal eddy viscosity, α is a constant to define a dry elements ($\alpha = 0$) and wet elements ($\alpha = 1$) (Le et al., 2020a), h is the bathymetry, g = 9.81 m/s2 is the gravitational acceleration, Cd is the bulk drag coefficient, τ_{wind} is the wind stress and ∇P_{atm} is the atmospheric pressure gradient. In order to run the hydrodynamic model, we define a computational domain that covers both the river and the ocean part, create an unstructured mesh to cover the domain, with a resolution of 50 m over the riverbanks, 400 m over the coast near the river mouth, 1 km over the rest of the coastline, and 5 km over the offshore (Fig. 2), set bathymetry and bulkdrag coefficient, imposed some forcings, and define the boundary conditions. More detail about the hydrodynamic model setup can be found in our previous work (Sampurno et al., 2021).

The hydrodynamic model simulation are forced by wind and atmospheric pressure from ECMWF (Hersbach et al., 2020), and tides from TPXO (Egbert and Erofeeva, 2002). As upstream boundary conditions, we imposed discharge from the Kapuas River and the Landak River. We also imposed runoff, which is obtained by converting rainfall over the Kapuas Kecil River catchment area as an inlet water flux at some channels entering the domain. The runoff of every channel was calculated from rainfall data using SWAT+ (Bieger et al., 2017), whose processes involved the pressure, the humidity, and other weather

https://doi.org/10.5194/npg-2021-36
Preprint. Discussion started: 4 January 2022


Nonlinear Processes in Geophysics Discussions

parameters as input. After a model simulate for a river scenario, we extracted the predicted water level at the observational point in Pontianak, and compared it with observational data to evaluate the model performance.

2.3 Flood scenarios

To simulate flood events, we considered 6000 scenarios with oceanic, atmospheric and river forcings (see Table 1). To produce the scenarios, we ran the hydrodynamic model for ten months. Then, we merged the output of all scenarios as a single dataset to train the machine learning model. Using these steps, we tried to encompass all possible flood events resulting from the any combination of the external forcings.

2.4 Machine learning model

2.4.1 Dependent and predictor variables

To develop machine learning models, we used the river water level at Pontianak as the dependent variable. Then, we considered atmospheric, oceanic, and riverine variables as the predictors (independent variables) of the water level in the city. Atmospheric variables include average and maximum wind speed, wind direction, precipitation, and average atmospheric pressure. Oceanic variables include tide at the river mouth, while the riverine variables include the Kapuas River and the Landak River discharges. To evaluate the impact of each predictor before the flood event, we also imposed the prior state (one and two hours before) of these parameters (see Table 2). The datasets were recorded hourly and combined the SLIM output (used in the training phase) and observational data (used in the testing phase).

atistic tool that can measure the degree of relatedness between variables in a dataset. The greater the MI value between two variables, the stronger its relatedness, regardless of how nonlinear its dependency (Kinney and Atwal, 2014). MI between two variables (X and Y) is obtained from Choi et al., (2020):

$$MI = \sum_{x \in X} \sum_{y \in Y} p(x,y)\, log\left(\frac{p(x,y)}{p(x)-p(y)}\right) \quad (3)$$

where $p(x,y)$ is the joint probability distribution.

2.4.2 Machine learning algorithm

Here, we consider three different machine learning algorithms, i.e., random forest (RF), multiple linear regression (MLR), and support vector machine (SVM). RF is a supervised learning algorithm that operates by constructing many decision trees at training time (Breiman, 2001). The algorithm can be implemented for classification or regression. To generate the ultimate output, the model aggregates its multiple decision trees outcomes, which are called sub-sample outcomes (Han et al., 2012). The technique was enhanced by combining bootstrap in its aggregating processes (Breiman, 2001). Using this strategy, the algorithm became an effective tool for classification and regression. In this study, the RF algorithm was obtained from the R randomForest library (Liaw and Wiener, 2002).

https://doi.org/10.5194/npg-2021-36
Preprint. Discussion started: 4 January 2022


Nonlinear Processes in Geophysics Discussions

MLR is an ML technique that aims to fit the linear relationship between input features and the target (observed data) using the least-squared approach (James et al., 2013). In the least-squared approach, the best relationship model will be obtained by minimizing the sum of the squared distance between the calculated values (as model outputs) and the target values (James et al., 2013). This algorithm is the most straightforward approach in machine learning models and generally used as the baseline method. The MLR algorithm which was implemented in this study is obtained from the R RWeka library (Hornik et al., 2008).

SVM is an ML algorithm that works based on statistical learning frameworks (Gholami and Fakhari, 2017). This method is robust for modeling a complex non-linear relationship. It amounts to choosing a kernel function that transforms the input features into a high-dimensional space to tackle the complexity. As a result, the process will transform the non-linear relationship of the input features to become linear. Finally, linear regression is carried out to obtain the ultimate output. Compared with the other algorithm, SVM needs less computational resources due to its capability to be trained only by a few features (Gholami and Fakhari, 2017). Previously, SVM was only implemented for classification purposes, yet, it has also been implemented for regression purposes after some enhancement. The SVM algorithm which we implement in this study is obtained from the R MARSSVRhybrid library (MARSSVRhybrid: MARS SVR Hybrid, 2021).

2.4.3. Metrics for machine learning model performance evaluation

Here, we use the Nash–Sutcliffe efficiency (NSE) measure to evaluate the models' performance. NSE is used to assess the performance of the machine learning models in producing the predicted water level. A perfect model corresponds to NSE = 1, while a model that has the same predictive skill as the mean of the observed data represents by NSE = 0. If NSE<0 implies that the mean value of observed data is predicting better than the model does. The closer NSE value to 1, the better the predictive skill of the model. The NSE coefficient is calculated as follows:

$$NSE = 1 - \frac{\sum_{t=1}^{T}\{\mathrm{H}_m^t - H_0^t\}^2}{\sum_{t=1}^{T}\left(H_0^t - \overline{\mathrm{H}_0}\right)^2} \quad (4)$$

where H_m^t represented the water level model at time t, H_0^t represented the observed water level at the same time, and $\overline{\mathrm{H}_0}$ is the mean of observed water level.

Root Mean Square Errors (RMSE) of peaks between predicted water level and observation during the flood events are also used as additional performance indicator. RMSE is used to represent the model's ability to predict flood events. The RMSE between the model outputs and the observations is calculated by:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(x_i - y_i)^2}{N}} \quad (5)$$

where x_i is the water level as the model's output at the i-th peaks, and y_i is the observed water level at the same time. N is the number of the total peaks data.

https://doi.org/10.5194/npg-2021-36
Preprint. Discussion started: 4 January 2022


Nonlinear Processes in Geophysics Discussions

3 Results

We ran a simulation for January 2019 and compared the simulated water elevation with observations in Pontianak. The model errors correspond to an NSE of 0.87 and an RMSE of 0.12 m, which are deemed sufficiently small to consider model outputs as a good proxy of the real system (Fig. 3).

We then simulated several inundation scenarios to produce datasets used to train the machine learning model. The simulations were run for ten months. Next, we extracted 6000 predicted water levels at Pontianak with their associated input dataset. Based on the Pontianak Maritime Meteorological Station report, floods occurred over the city when the water level exceeds 2.5 m. We therefore set that value as the threshold to define a flood event. It can be seen that several scenarios correspond to a water elevation greater than 2.5 m (Fig. 4).

All predictors considered in the machine learning model have mutual information (MI) coefficient greater than zero, which means that all predictor variables impact the river water level in Pontianak (Fig. 5). The relationship between these predictors and the water level could be linear or non-linear (as MI captured both types of relation). Here we found that the tidal elevations in the estuary (X1, X2, and X3) have the most decisive impact on the river water level in the city (MI > 0.5), while tidal elevation observed one hour before (X2) is the strongest one. Next, the wind speed (max and average), the discharges (from both the Kapuas and the Landak river), and the pressure have a moderate relatedness. Meanwhile, the wind direction and the rainfall have only a weak relatedness (MI < 0.1), which means that both parameters have no a significant impact.

Three machine learning models with different algorithms have been built to represent the relationship between dependent (hourly water level modelled with SLIM) and independent parameters (tide, discharges and weather parameters). Models were trained based on 6000 scenarios of the SLIM output data. During the training phase, all NSE coefficients are greater than 0.75, which means that all algorithms perform very well during the training phase (Fig. 6). The most accurate algorithm is the RF followed by the SVM and the MLR. We thus know that all tested machine learning algorithms are promising and need to be evaluated in the next testing phase to select the best one.

In order to evaluate the performance of each model in predicting real flooding events, we tested the machine learning models on the observational data. Testing data was selected during the high discharge season for three different months during which inundations occurred (Dec 2018, Jan 2020, Jan 2021). Fig. 7 shows the predicted water levels, produced by each proposed algorithm, compare to the observational data. The accuracy of models to predict flooding events, marked by points in Fig. 7, is evaluated in this phase.

Even though all algorithms perform very well during the training phase, the performances are different during the testing phases (Fig. 8 and Table 3). The RF accurately performs for three different testing phases. Its NSE values range from 0.61 to 0.72, which is a good performance. On the other hand, the MLR and the SVM algorithms succeed in the training phase and in the first and third testing phases but are less successful in the second testing phase. Their NSE on the second testing is only 0.39 and 0.40, respectively. Therefore, we know that the RF is the best machine learning algorithm in modeling water level dynamics for our test case.

https://doi.org/10.5194/npg-2021-36
Preprint. Discussion started: 4 January 2022


Nonlinear Processes in Geophysics Discussions

Regarding flood events prediction, the RF algorithm also performed better than the other algorithms. The algorithm could predict ten out of seventeen events with a RMSE of 0.11 m. On the other hand, MLR and SVM can only predict six and eight events with RMSE of 0.19 m and 0.18 m, respectively. All of these algorithms, however, also predicted four false-positive events, i.e., flood events that never occurred, during the testing phase.

4 Discussion

The two main issues that have been tackled in this study are data scarcity and low computational resources for building flood forecasting model in developing countries (Brocca et al., 2020; Singh et al., 2021). Here, we showed that using an approach that combines hydrodynamic and ML models is promising to obtain a reliable and robust model. We succeeded in building and evaluating ML models trained by the hydrodynamic model output; hence, they did not require extensive observational data in their training phase and did not need high computational cost in its implementation. Therefore, the proposed model is reliable for areas where observational data are scarce and computational resources are limited.

Since the proposed model can accurately forecast water levels, local water management agencies can rely on the model outputs for flood forecasting. Since the machine-learning did not required high computational resources, the limited computational resources will not prevent them from assessing and mitigating their compound flooding hazards. Using the model, they can re-assess their compound flood events and predict the future events. Moreover, once they have more observation data, they can fuse the data to re-adjust the proposed model or to build a more robust one (Muñoz et al., 2021).

Next, we found that the RF algorithm is the best ML algorithm to assess compound flooding in our area of interest. In general, the performances of all tested ML algorithms for water level prediction are reasonable and acceptable. However, considering the NSE values in all testing phases, the number of flood events that are accurately predicted, and how close the predicted water level is during the events, the RF performs clearly better than other algorithms. The superiority of the RF algorithm in predicting water levels has also been shown in previous studies in the Upo Wetland (Choi et al., 2020) and the Poyang Lake (Li et al., 2016). Therefore, we proposed a machine learning model with the RF algorithm as the most appropriate model to be implemented in the study area.

In addition, we also found that the tidal elevation measured one-hour prior at the river mouth is the main parameter controlling the river water level at Pontianak. Even though the city of Pontianak is located 20 km from the river mouth, the tidal dynamics still strongly affects the river water level changes in the city. This result confirms those of a previous study, which revealed that the tide propagation on the Kapuas River still impacts the water level in Sanggau, a city located 285 km from the river mouth (Kästner et al., 2019).

However, the machine learning model proposed here also has some limitations. Since the rainfall impact on river water level is minor compared with the other parameters, the model could not optimally capture the urban flooding over the city due to excessive rainfall. Based on the field observation, a short inundation occurs in the city if excessive rain occurs for a few hours. This inundation could be due to the poor quality of the urban drainage system. Unfortunately, this phenomenon is not directly

https://doi.org/10.5194/npg-2021-36
Preprint. Discussion started: 4 January 2022

captured by the water level observation located within the river. The increase of the river water level due to the heavy rain is minor. Next, the model relies on the predicted input parameters such as weather parameters and river discharges to predict the future water level. Consequently, the more biased the predictors, the higher the uncertainty in water-level prediction.

Overall, our integrated approach can provide a model to assess compound flooding driven by the interaction of tide, storm surge from the oceanward, and high discharge from the river upstream. However, the model still cannot optimally assess flooding caused by heavy rain over the city due to the limitation of the chosen indicator capability to capture the events. Therefore, we will enhance the model capability in future studies by considering more indicators representing flooding events due to excessive rainfall. Moreover, we will also try to reduce the number of predictors to minimize the model output's uncertainty. Mean sea-level rise due to climate change will also be evaluated to broaden the model implementation and create better flood mitigation.

4 Conclusion

This study shows that an integrated approach between the hydrodynamic and the machine learning models successfully overcomes modeling river water-level and predicting compound flooding in a data-scarce environment with limited computational resources. Therefore, the approach is suitable for local water management agencies in developing countries whose encountered those issues. Regarding the implementation in Pontianak, we found that the machine learning model with the RF algorithm has the most accurate output compared to the other algorithms. In addition, the tidal elevation, measured one hour before the current time, is the main predictor for river water level modeling in the study area.

Author contribution

JS, VV, and EH conceptualized the research; JS and RA curated the data; JS, VV, and EH analyzed the data; JS wrote the manuscript draft; JS and EH reviewed and edited the manuscript.

Competing interests

The authors declare that they have no conflict of interest.

Acknowledgment

The PhD fellowship of Joko Sampurno is provided by Indonesia Endowment Fund for Education (LPDP) under Grant No. 201712220212183. Computational resources have been provided by the supercomputing facilities of the Université catholique de Louvain (CISM/UCL) and the Consortium des Équipements de Calcul Intensif en Fédération Wallonie Bruxelles (CÉCI)

245 funded by the Fond de la Recherche Scientifique de Belgique (F.R.S.-FNRS) under convention 2.5020.11 and by the Walloon Region.

References

Assem, H., Ghariba, S., Makrai, G., Johnston, P., Gill, L., and Pilla, F.: Urban Water Flow and Water Level Prediction Based on Deep Learning, Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), 10536
250 LNAI, 317–329, https://doi.org/10.1007/978-3-319-71273-4_26, 2017.

Bieger, K., Arnold, J. G., Rathjens, H., White, M. J., Bosch, D. D., Allen, P. M., Volk, M., and Srinivasan, R.: Introduction to SWAT+, a Completely Restructured Version of the Soil and Water Assessment Tool, J. Am. Water Resour. Assoc., 53, 115–130, https://doi.org/10.1111/1752-1688.12482, 2017.

Bilskie, M. V. and Hagen, S. C.: Defining Flood Zone Transitions in Low-Gradient Coastal Regions, Geophys. Res. Lett., 45,
255 2761–2770, https://doi.org/10.1002/2018GL077524, 2018.

Breiman, L.: Random forests, Mach. Learn., 45, 5–32, https://doi.org/10.1023/A:1010933404324, 2001.

Brocca, L., Massari, C., Pellarin, T., Filippucci, P., Ciabatta, L., Camici, S., Kerr, Y. H., and Fernández-Prieto, D.: River flow prediction in data scarce regions: soil moisture integrated satellite rainfall products outperform rain gauge observations in West Africa, Sci. Reports 2020 101, 10, 1–14, https://doi.org/10.1038/s41598-020-69343-x, 2020.

260 Chan, N. W.: Impacts of Disasters and Disaster Risk Management in Malaysia: The Case of Floods, Resil. Recover. Asian Disasters, 239–265, https://doi.org/10.1007/978-4-431-55022-8_12, 2015.

Chen, J. H. and Asch, S. M.: Machine Learning and Prediction in Medicine — Beyond the Peak of Inflated Expectations, N. Engl. J. Med., 376, 2507, https://doi.org/10.1056/NEJMP1702071, 2017.

Choi, C., Kim, J., Han, H., Han, D., and Kim, H. S.: Development of Water Level Prediction Models Using Machine Learning
265 in Wetlands: A Case Study of Upo Wetland in South Korea, 12, 93, https://doi.org/10.3390/W12010093, 2020.

Comer, J., Indiana Olbert, A., Nash, S., and Hartnett, M.: Development of high-resolution multi-scale modelling system for simulation of coastal-fluvial urban flooding, Nat. Hazards Earth Syst. Sci., 17, 205–224, https://doi.org/10.5194/NHESS-17-205-2017, 2017.

Costabile, P. and Macchione, F.: Enhancing river model set-up for 2-D dynamic flood modelling, Environ. Model. Softw., 67,
270 89–107, https://doi.org/10.1016/J.ENVSOFT.2015.01.009, 2015.

Costabile, P., Costanzo, C., and Macchione, F.: A storm event watershed model for surface runoff based on 2D fully dynamic wave equations, Hydrol. Process., 27, 554–569, https://doi.org/10.1002/HYP.9237, 2013.

MARSSVRhybrid: MARS SVR Hybrid: https://cran.r-project.org/web/packages/MARSSVRhybrid/index.html, last access: 23 September 2021.

275 Egbert, G. D. and Erofeeva, S. Y.: Efficient inverse modeling of barotropic ocean tides, J. Atmos. Ocean. Technol., 19, 183–204, 2002.

https://doi.org/10.5194/npg-2021-36
Preprint. Discussion started: 4 January 2022

Frys, C., Saint-Amand, A., Le Hénaff, M., Figueiredo, J., Kuba, A., Walker, B., Lambrechts, J., Vallaeys, V., Vincent, D., and Hanert, E.: Fine-Scale Coral Connectivity Pathways in the Florida Reef Tract: Implications for Conservation and Restoration, Front. Mar. Sci., 7, 312, https://doi.org/10.3389/FMARS.2020.00312/BIBTEX, 2020.

Gholami, R. and Fakhari, N.: Support Vector Machine: Principles, Parameters, and Applications, in: Handbook of Neural Computation, Elsevier Inc., 515–535, https://doi.org/10.1016/B978-0-12-811318-9.00027-2, 2017.

Goltenboth, F., Timotius, K. H., Milan, P. P., and Margraf, J.: Ecology of insular Southeast Asia: the Indonesian archipelago, Elsevier B.V., 2006.

Han, J., Kamber, M., and Pei, J.: Data Mining: Concepts and Techniques, Elsevier Inc., https://doi.org/10.1016/C2009-0-61819-5, 2012.

Hao, Z. and Singh, V. P.: Compound Events under Global Warming: A Dependence Perspective, J. Hydrol. Eng., 25, 03120001, https://doi.org/10.1061/(ASCE)HE.1943-5584.0001991, 2020.

Haq, M., Akhtar, M., Muhammad, S., Paras, S., and Rahmatullah, J.: Techniques of Remote Sensing and GIS for flood monitoring and damage assessment: A case study of Sindh province, Pakistan, Egypt. J. Remote Sens. Sp. Sci., 15, 135–141, https://doi.org/10.1016/J.EJRS.2012.07.002, 2012.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., Chiara, G., Dahlgren, P., Dee, D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hogan, R. J., Hólm, E., Janisková, M., Keeley, S., Laloyaux, P., Lopez, P., Lupu, C., Radnoti, G., Rosnay, P., Rozum, I., Vamborg, F., Villaume, S., and Thépaut, J.: The ERA5 global reanalysis, Q. J. R. Meteorol. Soc., 146, 1999–2049, https://doi.org/10.1002/qj.3803, 2020.

Hornik, K., Buchta, C., and Zeileis, A.: Open-source machine learning: R meets Weka, Comput. Stat., 24, 225–232, https://doi.org/10.1007/S00180-008-0119-7, 2008.

Jadmiko, S. D., Murdiyarso, D., and Faqih, A.: Climate Changes Projection for Land and Forest Fire Risk Assessment in West Kalimantan, IOP Conf. Ser. Earth Environ. Sci., 58, 012030, https://doi.org/10.1088/1755-1315/58/1/012030, 2017.

James, G., Witten, D., Hastie, T., and Tibshirani, R.: An Introduction to Statistical Learning with Applications in R, Springer, 2013.

Kabenge, M., Elaru, J., Wang, H., and Li, F.: Characterizing flood hazard risk in data-scarce areas, using a remote sensing and GIS-based flood hazard index, Nat. Hazards, 89, 1369–1387, https://doi.org/10.1007/S11069-017-3024-Y/FIGURES/5, 2017.

Karamouz, M., Zahmatkesh, Z., Goharian, E., and Nazif, S.: Combined Impact of Inland and Coastal Floods: Mapping Knowledge Base for Development of Planning Strategies, J. Water Resour. Plan. Manag., 141, 04014098, https://doi.org/10.1061/(ASCE)WR.1943-5452.0000497, 2014.

Kästner, K., Hoitink, A. J. F., Torfs, P. J. J. F., Deleersnijder, E., and Ningsih, N. S.: Propagation of tides along a river with a sloping bed, J. Fluid Mech., 872, 39–73, https://doi.org/10.1017/JFM.2019.331, 2019.

Kinney, J. B. and Atwal, G. S.: Equitability, mutual information, and the maximal information coefficient, Proc. Natl. Acad.

Sci. U. S. A., 111, 3354–3359, https://doi.org/10.1073/pnas.1309933111, 2014.

Le, H.-A., Lambrechts, J., Ortleb, S., Gratiot, N., Deleersnijder, E., and Soares-Frazão, S.: An implicit wetting-drying algorithm for the discontinuous Galerkin method: application to the Tonle Sap, Mekong River Basin, Environ. Fluid Mech., 20, 923–951, https://doi.org/10.1007/s10652-019-09732-7, 2020a.

Le, H. A., Gratiot, N., Santini, W., Ribolzi, O., Tran, D., Meriaux, X., Deleersnijder, E., and Soares-Frazão, S.: Suspended sediment properties in the Lower Mekong River, from fluvial to estuarine environments, Estuar. Coast. Shelf Sci., 233, 106522, https://doi.org/10.1016/J.ECSS.2019.106522, 2020b.

Li, B., Yang, G., Wan, R., Dai, X., and Zhang, Y.: Comparison of random forests and other statistical methods for the prediction of lake water level: a case study of the Poyang Lake in China, Hydrol. Res., 47, 69–83, https://doi.org/10.2166/NH.2016.264, 2016.

Liaw, A. and Wiener, M.: Classification and regression by randomForest, R News, 2, 18–22, 2002.

Mokkenstorm, L. C., van den Homberg, M. J. C., Winsemius, H., and Persson, A.: River Flood Detection Using Passive Microwave Remote Sensing in a Data-Scarce Environment: A Case Study for Two River Basins in Malawi, Front. Earth Sci., 9, 552, https://doi.org/10.3389/FEART.2021.670997/BIBTEX, 2021.

Mosavi, A., Ozturk, P., and Chau, K.: Flood Prediction Using Machine Learning Models: Literature Review, Water 2018, Vol. 10, Page 1536, 10, 1536, https://doi.org/10.3390/W10111536, 2018.

Muñoz, D. F., Muñoz, P., Moftakhari, H., and Moradkhani, H.: From local to regional compound flood mapping with deep learning and data fusion techniques, Sci. Total Environ., 782, 146927, https://doi.org/10.1016/J.SCITOTENV.2021.146927, 2021.

Naqa, I. El, Ruan, D., Valdes, G., Dekker, A., McNutt, T., Ge, Y., Wu, Q. J., Oh, J. H., Thor, M., Smith, W., Rao, A., Fuller, C., Xiao, Y., Manion, F., Schipper, M., Mayo, C., Moran, J. M., and Haken, R. Ten: Machine learning and modeling: Data, validation, communication challenges, Med. Phys., 45, e834–e840, https://doi.org/10.1002/MP.12811, 2018.

Nayak, P. C., Sudheer, K. P., Rangan, D. M., and Ramasastri, K. S.: Short-term flood forecasting with a neurofuzzy model, Water Resour. Res., 41, 1–16, https://doi.org/10.1029/2004WR003562, 2005.

Planet dump retrieved from https://planet.osm.org: https://www.openstreetmap.org, last access: 20 October 2020.

Sampurno, J., Vallaeys, V., Ardianto, R., and Hanert, E.: Modeling interactions between tides, storm surges, and river discharges in the Kapuas River delta, Biogeosciences Discuss. [preprint], https//doi.org/10.5194/bg-2021-273, Rev. 2021., https://doi.org/10.5194/BG-2021-273, 2021.

Santiago-Collazo, F. L., Bilskie, M. V., and Hagen, S. C.: A comprehensive review of compound inundation models in low-gradient coastal watersheds, Environ. Model. Softw., 119, 166–181, https://doi.org/10.1016/J.ENVSOFT.2019.06.002, 2019.

Semedi, P.: Palm Oil Wealth and Rumour Panics in West Kalimantan, 41, 233–252, https://doi.org/10.1080/08039410.2014.901240, 2014.

Singh, R. K., Soni, A., Kumar, S., Pasupuleti, S., and Govind, V.: Zonation of flood prone areas by an integrated framework of a hydrodynamic model and ANN, 21, 80–97, https://doi.org/10.2166/WS.2020.252, 2021.

https://doi.org/10.5194/npg-2021-36
Preprint. Discussion started: 4 January 2022

345 Tucci, C. E. M. and Villanueva, A. O. N.: Flood control measures in União da Vitoria and Porto União: structural vs. non-structural measures, 1, 177–182, https://doi.org/10.1016/S1462-0758(00)00012-1, 1999.

Vallaeys, V., Kärnä, T., Delandmeter, P., Lambrechts, J., Baptista, A. M., Deleersnijder, E., and Hanert, E.: Discontinuous Galerkin modeling of the Columbia River's coupled estuary-plume dynamics, Ocean Model., 124, 111–124, https://doi.org/10.1016/j.ocemod.2018.02.004, 2018.

350 Vallaeys, V., Lambrechts, J., Delandmeter, P., Pätsch, J., Spitzy, A., Hanert, E., and Deleersnijder, E.: Understanding the circulation in the deep, micro-tidal and strongly stratified Congo River estuary, Ocean Model., 167, 101890, https://doi.org/10.1016/J.OCEMOD.2021.101890, 2021.

Wang, Q. and Wang, S.: Machine Learning-Based Water Level Prediction in Lake Erie, Water 2020, Vol. 12, Page 2654, 12, 2654, https://doi.org/10.3390/W12102654, 2020.

355

https://doi.org/10.5194/npg-2021-36
Preprint. Discussion started: 4 January 2022

Nonlinear Processes in Geophysics
Discussions
Open Access
EGU

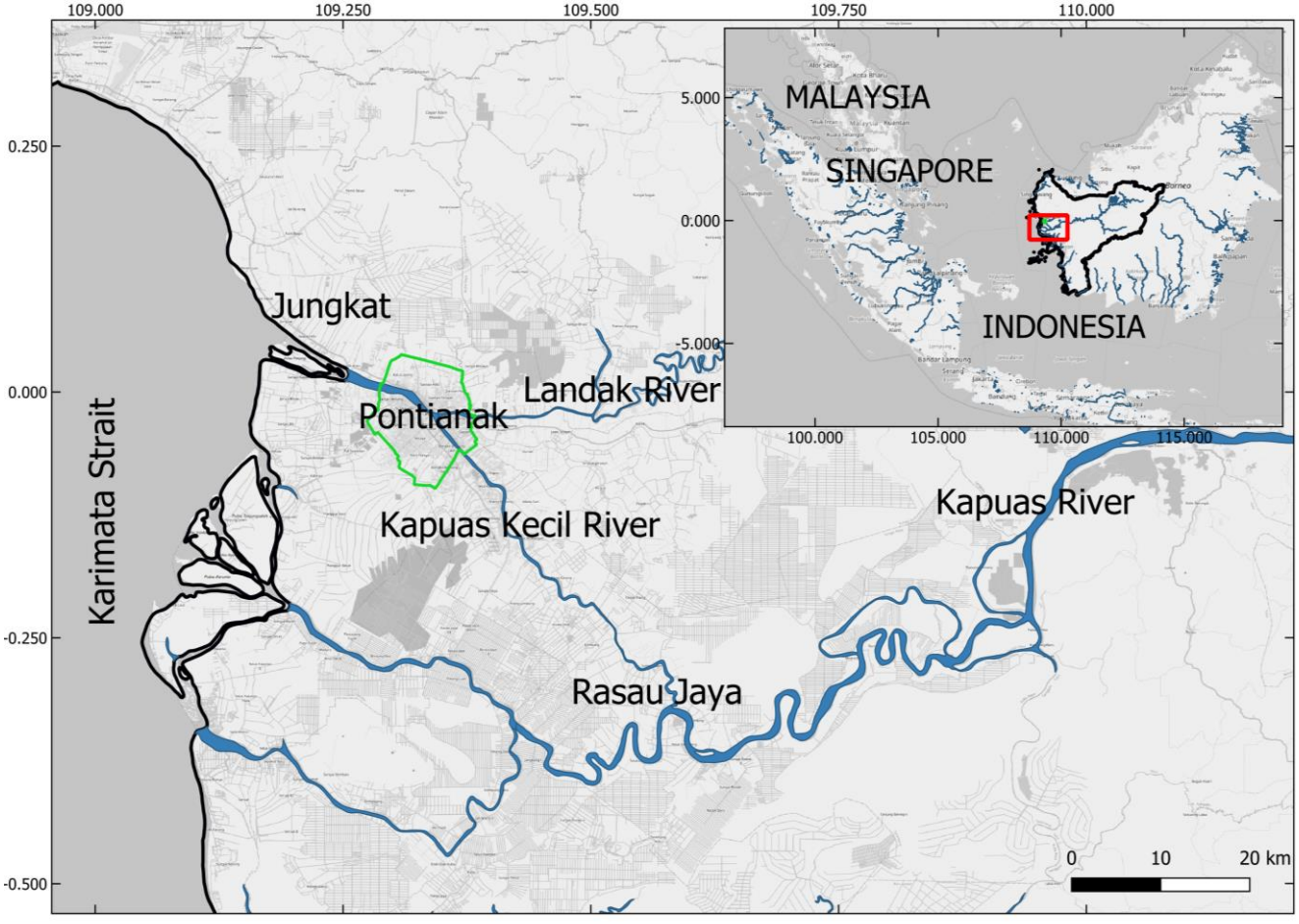

Figure 1: The region of interest where the city of Pontianak is represented by the green enclosed perimeter. Background map retrieved from (Planet dump retrieved from https://planet.osm.org, 2020)**. © OpenStreetMap contributors 2017. Distributed under the Open Data Commons Open Database License (ODbL) v1.0.**

https://doi.org/10.5194/npg-2021-36
Preprint. Discussion started: 4 January 2022

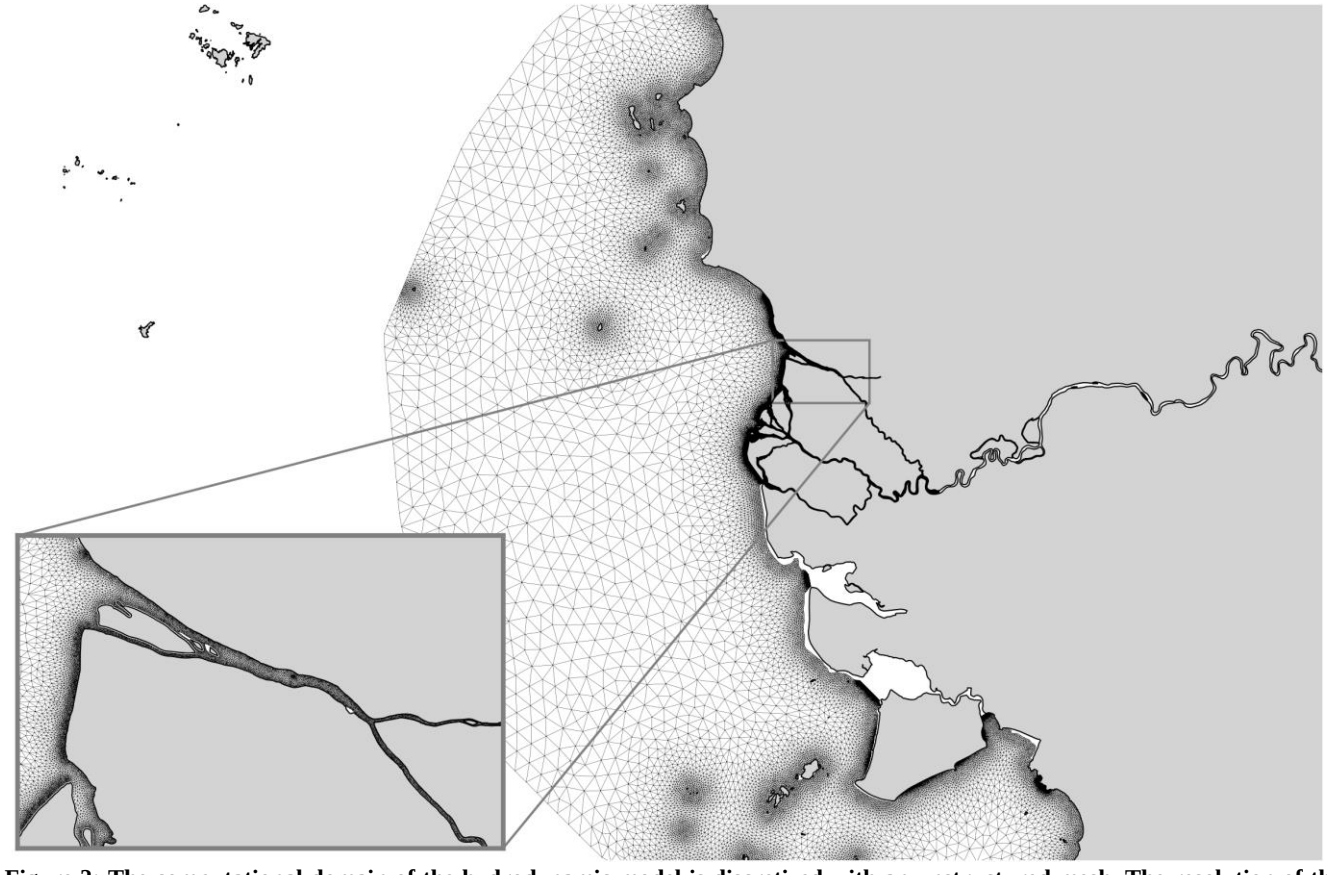

Figure 2: The computational domain of the hydrodynamic model is discretized with an unstructured mesh. The resolution of the mesh is set to 50 m along the riverbanks, 400 m along the coast near the estuary, 1 km over the rest of the coastline, and 5 km offshore. Background map retrieved from (OpenStreetMap contibutors, 2017). © OpenStreetMap contributors 2017. Distributed under the Open Data Commons Open Database License (ODbL) v1.0.

https://doi.org/10.5194/npg-2021-36
Preprint. Discussion started: 4 January 2022

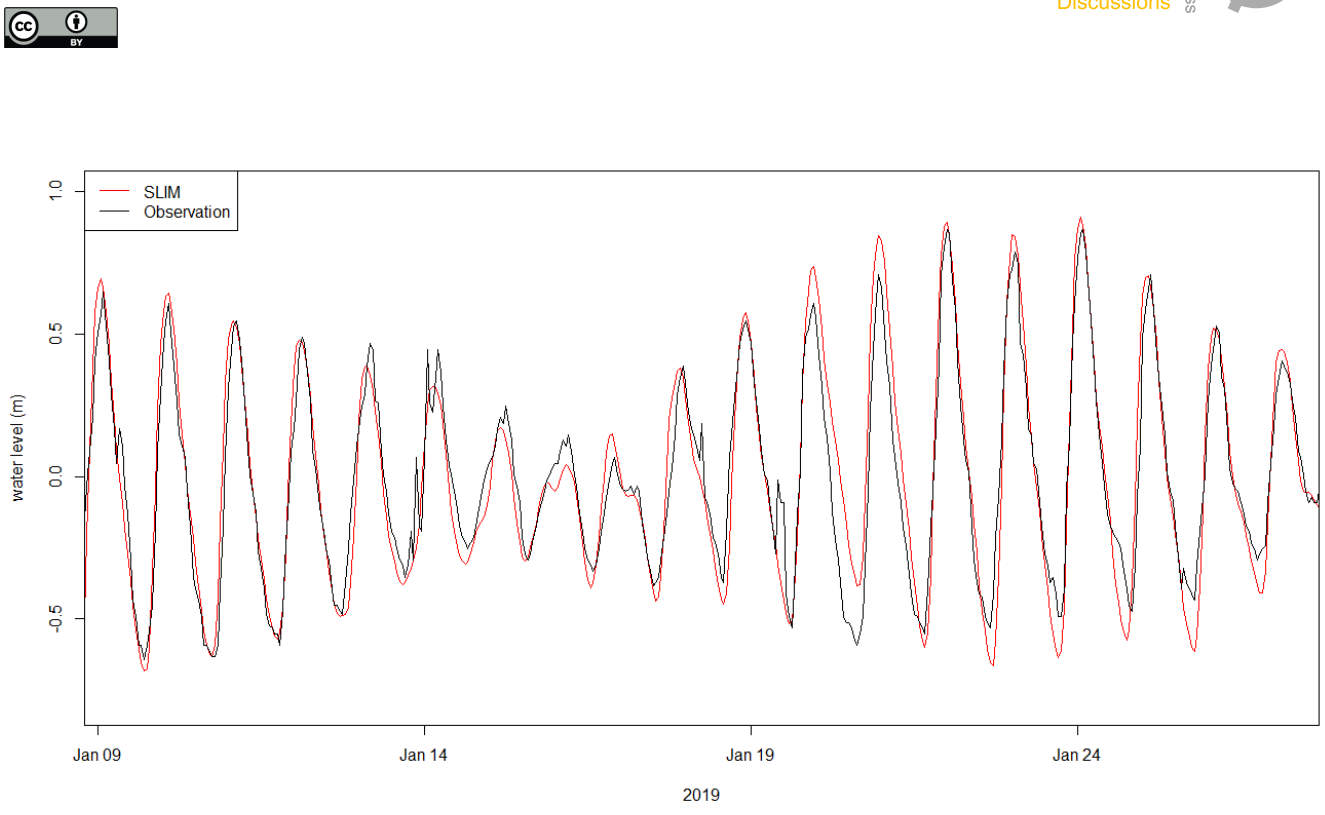

Figure 3: SLIM model output validation with respect to observational data at Pontianak in January 2019, with NSE = 0.87 and RMSE = 0.12 m, indicates that the model has satisfactory performance.

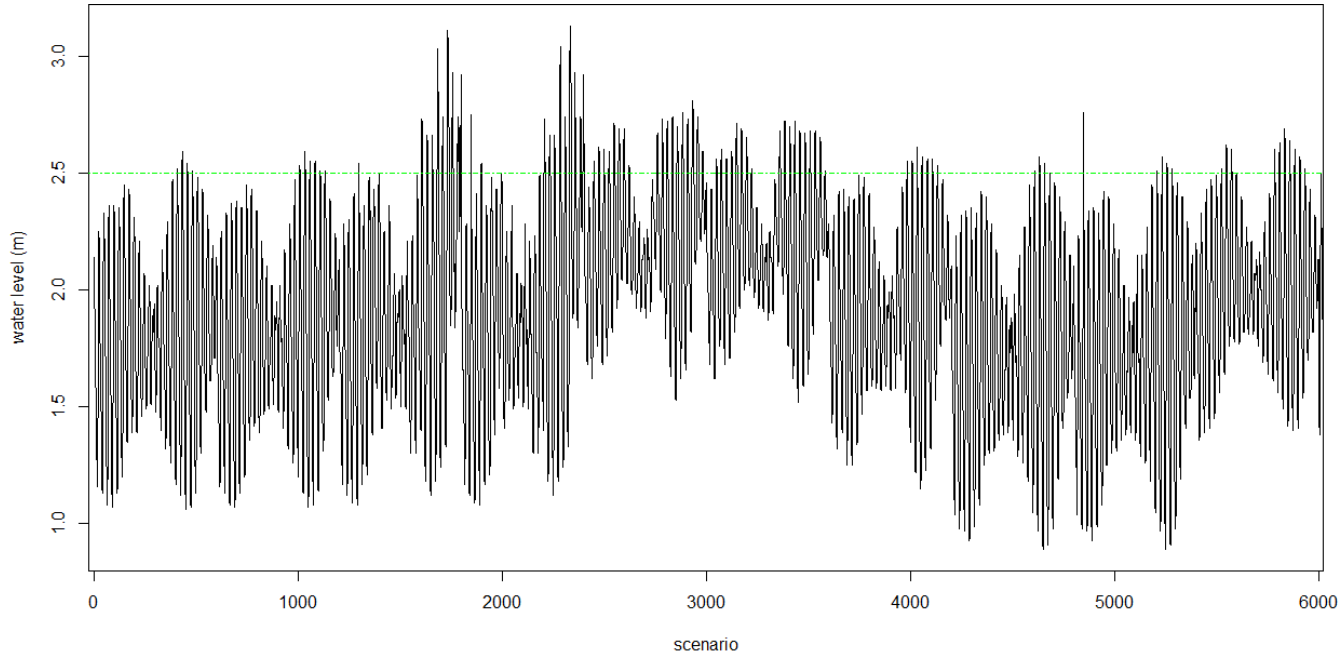

Figure 4: The Kapuas Kecil River's water level in Pontianak, obtained from the hydrodynamic model. The green dash line is the threshold above which the water starts to overflow the riverbanks in Pontianak.

https://doi.org/10.5194/npg-2021-36
Preprint. Discussion started: 4 January 2022

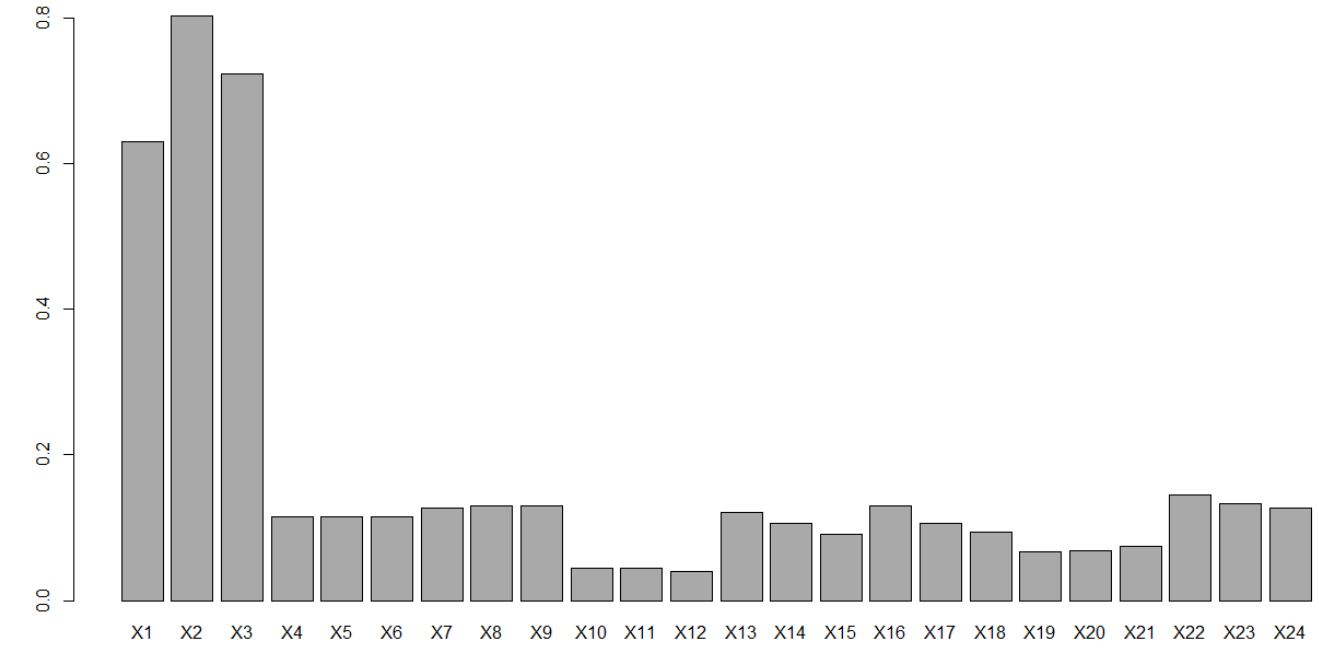

Figure 5: Mutual information of all independent variables to hourly water level dynamics in 3 months of observational data.

380

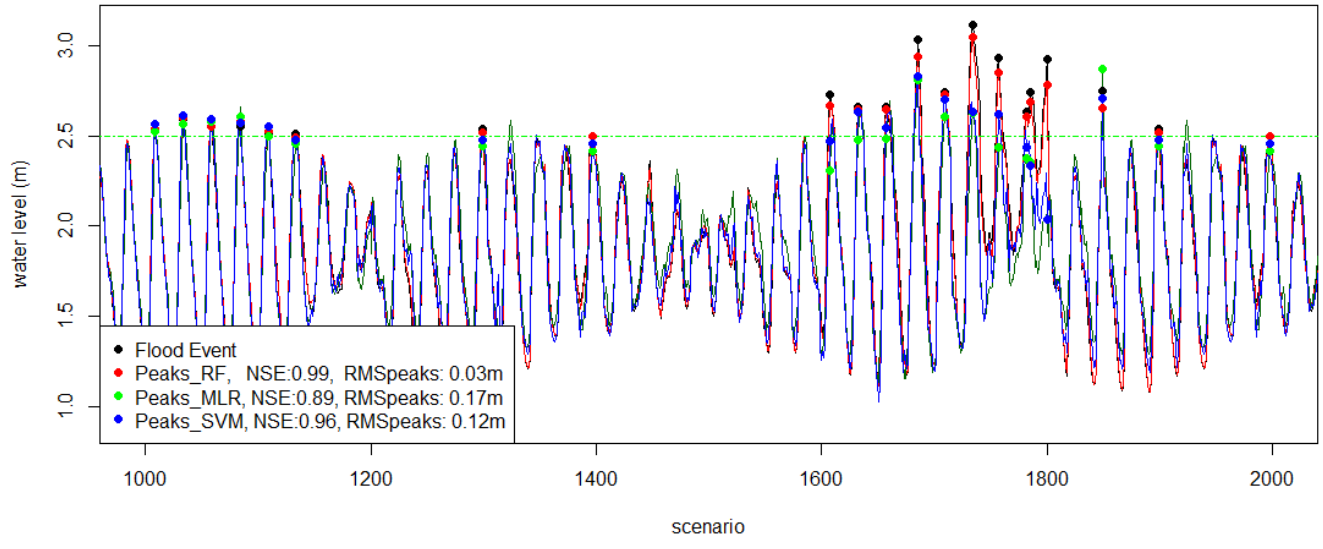

Figure 6: Comparison of predicted and observed hourly water levels of the training data.

https://doi.org/10.5194/npg-2021-36
Preprint. Discussion started: 4 January 2022

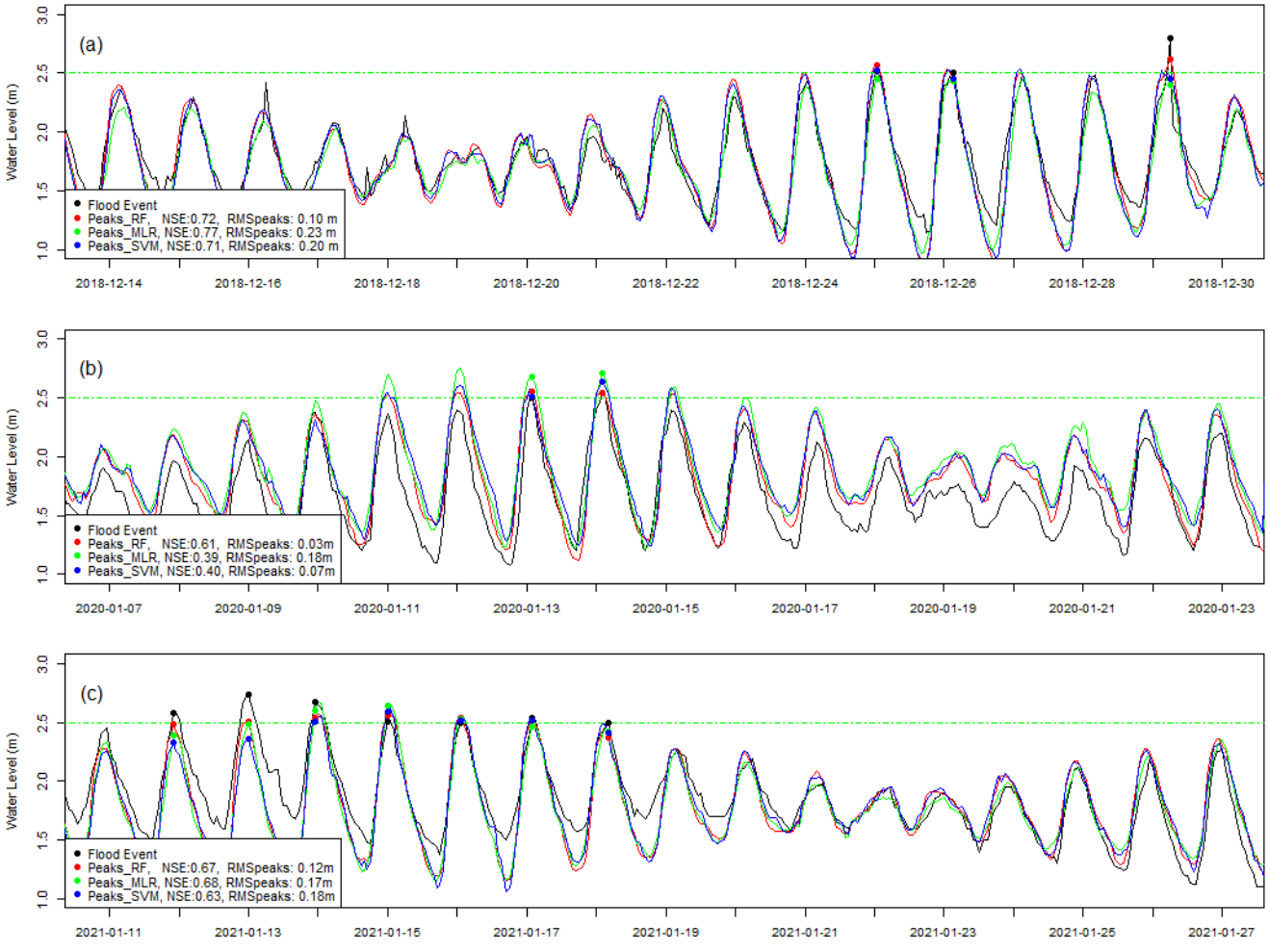

Figure 7: Comparison of predicted hourly water levels models and measured hourly water levels for the testing data on: (a)December 2018, (b) January 2020, and (c) January 2021

https://doi.org/10.5194/npg-2021-36
Preprint. Discussion started: 4 January 2022

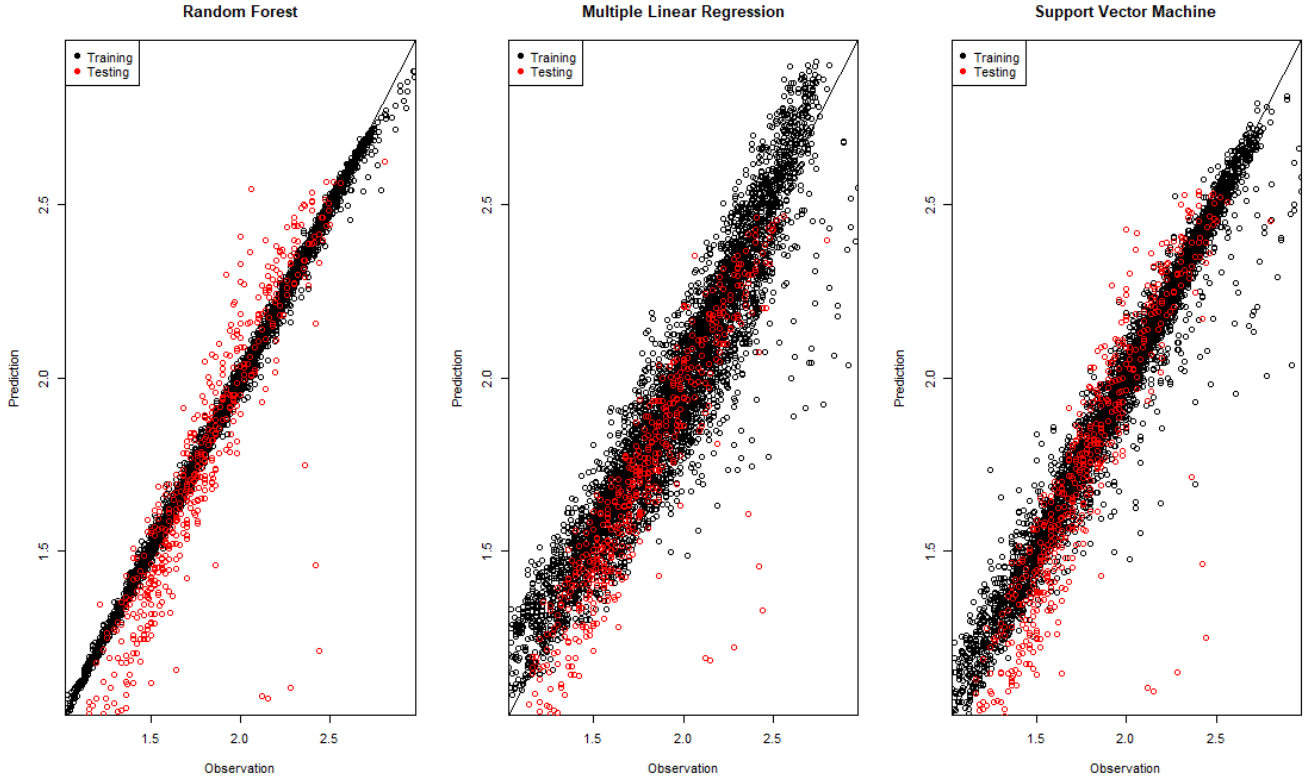

Figure 8: Performance of models in training and first testing phase

https://doi.org/10.5194/npg-2021-36
Preprint. Discussion started: 4 January 2022

Nonlinear Processes in Geophysics
Discussions
Open Access
EGU

400

Table 1. Scenarios used to force the process-based hydrodynamic model

n-scenario	Wind Speed (ms^{-1})	Wind Direction (°)	Pressure (kPa)	Discharge Kapuas (m^3s^{-1})	Discharge Landak (m^3s^{-1})	Rainfall (mm)
1-600	2 - 8	0 – 360	100.5 - 101.5	6×10^3	600	0
601-1200	4-16	0 – 360	100.5 - 101.5	6×10^3	600	0
1201-1800	8-32	0 – 360	100.5 - 101.5	6×10^3	600	0
1801-2400	2 – 8	0 – 360	100.5 - 101.5	10^4 - 1.5×10^4	600	0
2401-3000	2 – 8	0 – 360	100.5 - 101.5	6×10^3	800-2100	0
3001-3600	2 – 8	0 – 360	100.5 - 101.5	10^4 - 1.5×10^4	800-2100	0
3601-4200	5 – 20	0 – 360	100.5 - 101.5	10^4 - 1.5×10^4	800-2100	0
4201-4800	2 – 8	0 – 360	100.5 - 101.5	$3.3 \times 10^3 – 5 \times 10^3$	250-700	0
4801-5400	2 – 8	0 – 360	100.5 - 101.5	$3.3 \times 10^3 – 5 \times 10^3$	250-700	0 – 150
5401-6000	8-32	0 – 360	100.5 - 101.5	6×10^3	600	0 – 150

Table 2. The variables which used as the predictors in this study.

Code	Variable Description
X1	Tidal Elevation at Kapuas Kecil river mouth (m)
X2	Tidal Elevation at Kapuas Kecil river mouth 1 hour before (m)
X3	Tidal Elevation at Kapuas Kecil river mouth 2 hours before (m)
X4	Hourly Discharge of the Kapuas River at Rasau Jaya in time (m^3s^{-1})
X5	Hourly Discharge of the Kapuas River at Rasau Jaya 1 hour before (m^3s^{-1})
X6	Hourly Discharge of the Kapuas River at Rasau Jaya 2 hours before (m^3s^{-1})
X7	Hourly Discharge of the Landak River at Kuala Mandor in time (m^3s^{-1})
X8	Hourly Discharge of the Landak River at Kuala Mandor 1 hour before (m^3s^{-1})
X9	Hourly Discharge of the Landak River at Kuala Mandor 2 hours before (m^3s^{-1})
X10	Hourly Precipitation at the time (mm)
X11	Hourly Precipitation one hour before (mm)
X12	Hourly Precipitation two hours before (mm)
X13	Hourly Average Wind Speed at the time (ms^{-1})
X14	Hourly Average Wind Speed one hour before (ms^{-1})
X15	Hourly Average Wind Speed two hours before (ms^{-1})
X16	Hourly Maximum Instantaneous Wind Speed at the time (ms^{-1})
X17	Hourly Maximum Instantaneous Wind Speed one hour before (ms^{-1})
X18	Hourly Maximum Instantaneous Wind Speed two hours before (ms^{-1})
X19	Hourly Average Wind Direction at the time (degree, in the range: 0 - 360)
X20	Hourly Average Wind Direction one hours before (degree, in the range: 0 - 360)
X21	Hourly Average Wind Direction two hours before (degree, in the range: 0 - 360)
X22	Hourly Atmospheric Pressure at the time (millibars)
X23	Hourly Atmospheric Pressure one hour before (millibars)
X24	Hourly Atmospheric Pressure two hours before (millibars)

https://doi.org/10.5194/npg-2021-36
Preprint. Discussion started: 4 January 2022

Nonlinear Processes in Geophysics Discussions
Open Access
EGU

405 **Table 3.** Performance evaluation of the three machine learning algorithms

Goodness of Fit	RF	SVM	MLR
NSE Training	0.99	0.96	0.89
NSE Testing 1	0.72	0.71	0.77
NSE Testing 2	0.61	0.40	0.39
NSE Testing 3	0.67	0.63	0.68
Total of flood predicted (out of 17 events)	10	8	6
Percent of flood predicted	59%	47%	35%
RMSE during flood events (m)	0.11	0.18	0.19
False Positive (event)	4	4	4