20th International Congress on Modelling and Simulation, Adelaide, Australia, 1–6 December 2013
www.mssanz.org.au/modsim2013

Numerical weather models as virtual sensors to data-driven rainfall forecasts in urban catchments

L. Cozzi [a], **S. Galelli** [b], **A. Castelletti** [a], **S. Jolivet** [c]

[a] *Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, Milano, Italy*
[b] *Pillar of Engineering System & Design, Singapore University of Technology and Design, Singapore*
[c] *Singapore Delft-Water Alliance, National University of Singapore, Singapore*
Email: stefano_galelli@sutd.edu.sg

Abstract: Weather and rainfall forecasts play a key role in operational urban water management, since they allow anticipating the effects of storms and flash floods and initiating alarms in a timely manner. Short-term rainfall forecasts are commonly based on radar nowcasting techniques, which provide high-resolution forecasts limited to a temporal horizon of few hours. Numerical weather models can extend this horizon further in time, but their coarser spatial resolution is often a limit when working on small urban catchments.

In this study, with the purpose of providing long-term and reliable rainfall forecasts in a small urban area, we investigate the use of numerical weather models as *virtual sensors*. The Weather Research and Forecast (WRF) model, fed with real-time Global Forecast System (GFS) data, is implemented for Singapore spatial domain and used to get a time-continuous and spatially-distributed monitoring of the atmospheric processes. The model state variables (e.g. humidity, air temperature, heat exchange etc.) are then processed by an automatic input variable selection algorithm to single out the most relevant variables used by a data-driven model to yield rainfall forecasts at the catchment scale. In this work, we explore different lead times (up to 12 hours) to evaluate the reliability of this approach, as well as different meteorological seasons.

A comparison against the forecasts issued by WRF shows that the prediction accuracy of the input selection-based data-driven models can be improved, especially for long-term predictions (up to 12 hours). Results show that for short lead times (up to 3 hours) the heat flux is the most relevant driver, while for longer lead times a combination of drivers, such as wind and temperature, is selected. Also, such combination varies with the meteorological season. This techniques can thus be adopted to improve the accuracy of rainfall forecasts, although it must be noted that the overall accuracy is still influenced by the underlying numerical model. Indeed, if the numerical weather model does not adequately represent some events (e.g. small-scale convective storms), the selected variables cannot be successfully adopted to yield a rainfall forecast.

Further research will focus on two different aspects. First, the rainfall forecasts will be included in an operational framework, in order to provide long-term inflow predictions to Marina Reservoir and to assess their impact on the barrage operation. Second, the results here discussed will be further investigated, by running the input selection algorithm on a larger datasets and by considering different combinations of lead times and meteorological seasons.

Keywords: *Rainfall forecasts; Input variable selection; Data-driven models; Urban water systems*

1 INTRODUCTION

Urban water reservoirs, fed by drainage systems, are becoming an attractive solution to increase the availability of drinking water in large metropolitan areas. Even though they ensure a strategic service and long-term sustainability (Kristiana et al., 2011), these reservoirs pose a number of challenges from an operational perspective, mainly because of the short time of concentration and increased surface runoff characterizing urban catchments (Zoppou, 2001). Their operation can therefore benefit from the availability of long-term inflow predictions, which, on their turn, are driven by rainfall forecasts.

The different techniques developed to provide quantitatively reliable rainfall forecasts can be reorganized in two macro groups, namely radar nowcasting and numerical modeling. Nowcasting techniques evaluate the directions of clouds fronts by comparing different radar scans to extrapolate the position and intensity of rainfall events (Wilson et al., 1998). The term nowcasting underlines that the predictions are time- and space-specific, for prediction horizons generally not longer than few hours. Indeed, the predictive capability of nowcasting techniques rapidly decreases with the lead time, since they can not account for the development and decay of precipitation. These techniques are generally used to forecast rainfall events over small spatial domains (e.g. urban areas) because of their high resolution. Numerical models predict rainfall events by accounting for the evolution of the state of the atmosphere, represented by fluids dynamics and thermodynamics equations (Baer, 2000), and whose resolution requires significant computational efforts. The predictive capability of these models is strictly related to their spatial resolution: as discussed by Golding (2009), numerical models accurately resolve only meteorological features five times bigger than the model grid cells. Therefore, both techniques cannot be successfully applied to support the operation of urban water reservoirs. On one hand, nowcasting techniques provide rainfall forecasts at the resolution needed for urban catchments, but their application is limited to short-term horizons. On the other hand, numerical models can extend the prediction horizon further in time, but they cannot be implemented at a high spatial resolution because of their computational requests. Indeed, at the smallest scales their accuracy decreases and they thus fail in predicting the exact precipitation intensity.

In this study, with the purpose of providing long-term and reliable rainfall forecasts for a small urban catchment, we investigate the use of numerical weather models as *virtual sensors*. The Weather Research and Forecast (WRF) model, fed with real-time Global Forecast System (GFS) data, is implemented for Singapore domain and used to get a time-continuous and spatially-distributed monitoring of the main atmospheric processes. The model state variables (e.g. humidity, air temperature, heat exchange etc.) are then processed by an automatic input variable selection algorithm to single out the most relevant variables, used by a data-driven model to yield rainfall forecasts at the catchment scale. In this work, we explore different lead times (up to 12 hours) to evaluate the reliability of this approach. Results show that the proposed approach allows for a significant improvement in the prediction accuracy of the original WRF model.

The paper is structured as follows. Section 2 gives a short explanation of the case study and adopted methodology, Section 3 describes the data used in this work, while Section 4 discusses the experimental results. Finally, Section 5 highlights conclusions and further research needs.

2 MATERIAL AND METHODS

2.1 Study site

When Singapore became an independent state in 1965, water imported from Malaysia was its main source of drinking water, although during the years many efforts have been done to reach self-sufficiency. The strategy called *Four National Taps* specifies which sources must be taken into account (Xie, 2006): i) catchments management; ii) Water imported from Malaysia; iii) Wastewater treatment (NEWater); iv) Desalinated water. The core of the first tap is the idea of employing the largest catchment in Singapore, Marina Reservoir catchment, as a source of drinking water supply. For this reason, a reservoir was built in late 2008 by constructing a dam in the former Marina. The surrounding catchment has an approximated surface of 100 km^2, and produces a mean annual inflow of about 150 Mm^3. The time of concentration is about one hour, and this makes the reservoir very sensitive to flow peaks. A previous study demonstrated that the management of the reservoir can benefit from the use of 6-9 hours ahead inflow predictions based on rainfall forecasts (Galelli et al., 2012).

Being situated almost on the Equator, Singapore is characterized by a tropical climate. There are two monsoon and two inter-monsoon seasons. December is the wettest month of the year, with a monthly mean rainfall of 330 mm (Fong, 2012), and the total annual precipitation is about 2200 mm. In tropical environments physical

processes behind rainfall events are usually fast and characterised by a small spatial scale, increasing the accuracy that a model must have to reproduce them. Three types of rainfall events usually occur in Singapore: i) Thunderstorms, i.e. fast convective events due to the heat exchange between the surrounding ocean and the atmosphere; ii) Sumatra squall lines, convective events formed over the island of Sumatra and that reach Singapore at dawn; iii) Monsoons, long-lasting events due to the seasonal winds blowing in summer and winter seasons. Except for the northeast monsoon season, which is mainly characterized by monsoon events, any other season presents intense rainfall events, especially during the southwest monsoon season, when Sumatra squall lines are frequent (Fong, 2012).

2.2 Weather Research and Forecast (WRF) model

In this work, we used WRF model (version 3.3) to describe Singapore meteorological system. WRF is a local area model developed by a partnership between the National Center for Atmospheric Research (NCAR) and the National Oceanic and Atmospheric Administration (NOAA) (Skamarock et al. (2008), Wang et al. (2011)).

The spatial domain is composed of 30x60 grid cells with a resolution of 1 km^2, resulting in an area of 30x60 km^2 that includes the island of Singapore and the surrounding ocean. On the vertical profile 27 layers are considered, thus leading to a total of 48,600 computational cells. Every 6 hours WRF is fed by GFS data, used as boundary conditions to run a new simulation. GFS is a numerical weather prediction system developed by NOAA based on the primitive dynamical equations. It combines 6-hours forecasts with atmospheric observations to produce initial conditions.

The WRF model has a total of 41 state variables (per cell) divided in 6 main groups: heat, temperature, wind, humidity, pressure, and geopotential. The values of the state variables are spatially distributed according to the model resolution, resulting in more than $1.5 \cdot 10^6$ state variables, evaluated with a time step of 10 minutes.

2.3 Tree-based input variable selection

In order to understand which atmospheric variables modelled by WRF can be used by a data-driven model as predictors of rainfall events, the large-dimensional dataset produced by WRF is processed with an input variable selection algorithm. In particular, we resort to the Iterative Input variable Selection (IIS) algorithm (Galelli and Castelletti, 2013b), which scales well to large datasets and accounts for non-linear dependencies and redundancy between variables.

Given a sample dataset composed of N observations of the output variable y and n candidate inputs, the IIS algorithm solves the problem of input selection in three steps. Firstly, it ranks the n candidates according to a significant measure of relevance (i.e. the explained variance). Secondly, the first p ranked variables (with $1 \leqslant p \leqslant n$) are used to build p Single-Input-Single-Output (SISO) models to re-assess their relevance in explaining y. Then, the variable used in the best SISO model is added to the set X_y of selected variables. Thirdly, a Multi-Input-Single-Output (MISO) model is identified using the set X_y of selected variables, and its performance is evaluated. The model residuals are used to iterate the first two steps, until the performance of the MISO model does not significantly improve.

This procedure requires an algorithm for building a data-driven model in order to find a proper relation between candidate inputs (i.e. the weather variables) and output (i.e. measured rainfall). In this study, the IIS algorithm is combined with *Extremely Randomized Trees*, a non-parametric tree-based ensemble regression method proposed by Geurts et al. (2006). See Galelli and Castelletti (2013a) for a study of their application to hydrological problems.

3 Experimental set-up

The dataset used in this work consists of time series covering a 21 months period, from April 2009 until December 2010. The dataset is produced as follows. Every 6 hours a new WRF simulation is run, adopting GFS data as boundary conditions and calculating a new value for the state variables every 10 minutes. The repetition of these simulations over the 21-months period leads to a dataset of 92,160 observations for each state variable calculated by WRF. These are the input variables to be correlated with the rainfall measured over Marina catchment. This latter is obtained by interpolating with Thiessen polygons the data registered by different stations.

The performance of the IIS algorithm depends on the ratio between the number of observations and input variables: the lower is the ratio, the more difficult is the input selection problem. To reduce the number of input variables, we consider a smaller area of about 100 km^2 corresponding to Marina Reservoir catchment,

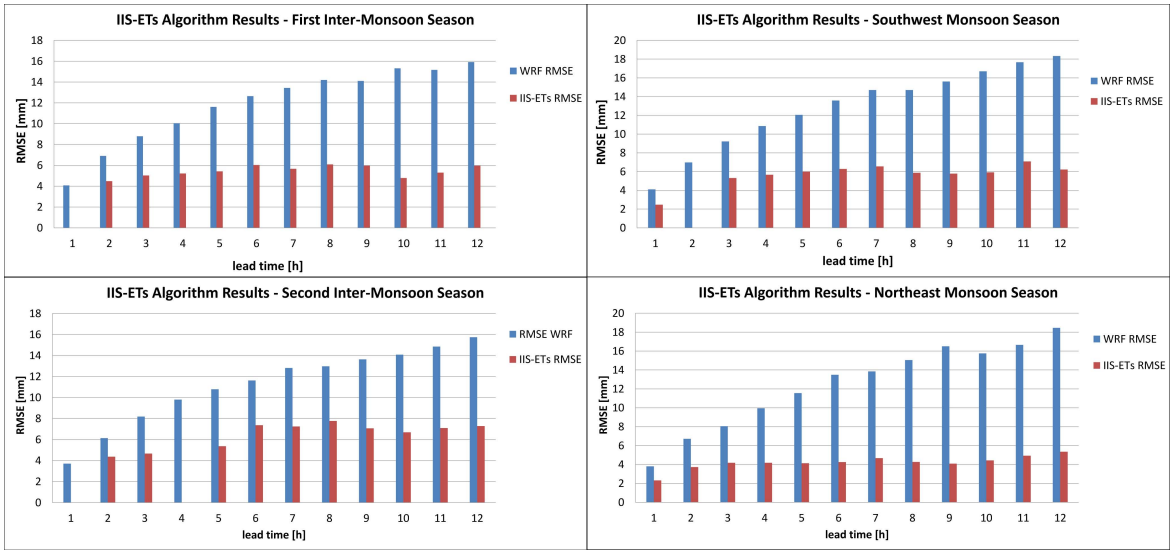

Figure 1. Predictive accuracy (in terms of Root Mean Squared Error, RMSE) of WRF and input selection-based data-driven models (IIS-ETs) over the four meteorological seasons.

over which WRF state variables are spatially averaged. As for the vertical profile, we consider two layers only, one in correspondence to the height of 1.5 km, and a pressure value of 850 mb, the other in correspondence to the height of 6 km, and pressure value of 150 mb. In this way the total number of input variables is reduced to 67. As for the temporal resolution, we decided to aggregate the modelled variables with a hourly time step. Finally, with the purpose of exploring different lead times and assessing the approach effectiveness over different meteorological seasons, 48 different experiments are defined as follows.

- *Lead time*. In this work we consider experiments with a prediction horizon comprised between 1 to 12 hours, with the output variable defined as the cumulated rainfall over the prediction horizon. Notice that for longer lead times this leads to output variable time series characterised by a lower variance, as different events registered during few hours may be cumulated and represented as a single event.
- *Meteorological seasons*. Singapore's climate is characterized by four meteorological seasons, with three different types of rainfall events. Each of them is triggered by different drivers, so we can in principle assume that the processes behind rainfall events are periodic and time-variant. Thus, in order to achieve the best results from the input selection experiments, it is not possible to include an entire year within a single dataset. Therefore, we chose to define a different experiment for each meteorological season.

4 Results

Figure 1 shows the results from the input variable selection experiments, and compares the predictive capability of WRF and the data-driven models based on the input selection exercises. It is evident that the adoption of the proposed approach allows for a strong improvement in the accuracy of the rainfall forecasts. Moreover, results show that the accuracy of the input selection-based data-driven models increases with the lead time. This is probably due to the variability of the model output: for large values of the lead time the measured rainfall is cumulated over an interval of some hours (up to 12) and its variance is lower, while for smaller values of the lead time the input selection results are affected by the high variability of the output variable.

The best results are obtained for the northeast monsoon season, with the value of RMSE never exceeding 6 mm, while the results for the remaining three seasons are less satisfactory. This is probably due to the type of modelled rainfall events: indeed, the most frequent rainfall events in the winter seasons are monsoons, which can be reproduced by the model because of their large spatial scale and slow dynamic. On the other hand, thunderstorms are small-scale events with fast dynamics, which cannot be fully captured by the model.

To understand which are the most relevant drivers to predict rainfall events, we gathered WRF state variables in six groups, namely heat, humidity, pressure, temperature, wind and geopotential, and reported the relative

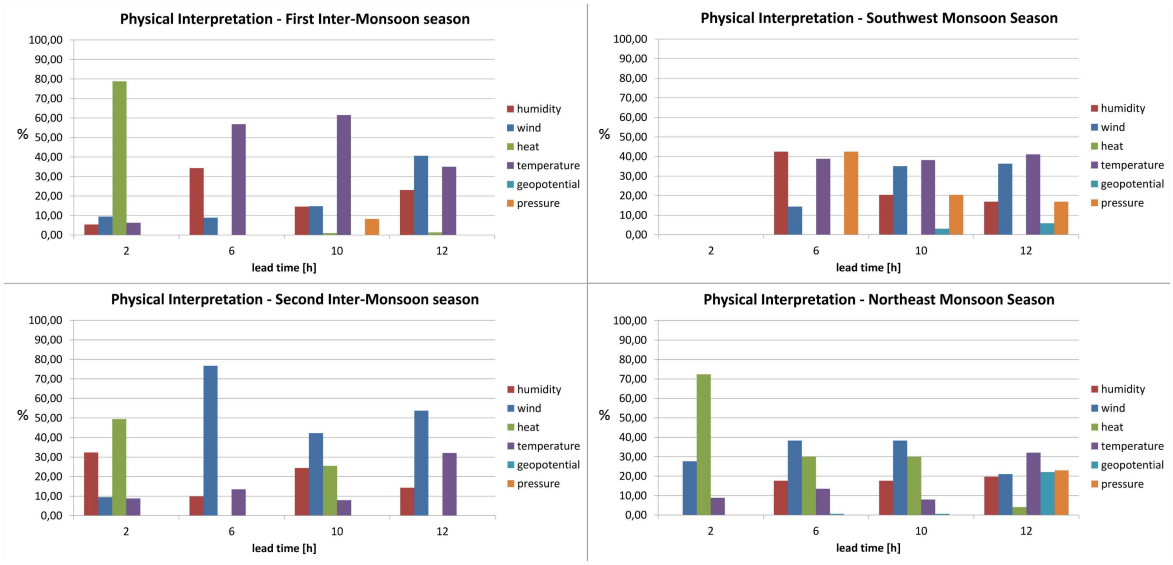

Figure 2. Relative importance of each group of variables, in terms of contribution to the model explained variance, for each meteorological season. The values are calculated for experiments with a lead time of 2, 6, 10 and 12 hours.

importance of each group in explaining the output variable for different temporal horizons. This analysis (see Figure 2) shows that heat is the most important driver for short-term predictions, probably because heat fluxes are the most important driver ruling convective storms. When using longer lead times, the dominant group varies with the considered season: for the two inter-monsoon seasons the relevant drivers are temperature and wind (first inter-monsoon season) or temperature (second inter-monsoon season), whereas for the two monsoon seasons all the variables have the same relative importance.

5 Conclusions

In this work we investigate the possibility of using the meteorological information described by WRF model as predictors to enhance the accuracy of data-driven rainfall forecasts. Through the use of an input variable selection algorithm, the most relevant variables among those provided by the numerical weather model are selected and used as regressors by a data-driven model. A comparison against the forecasts issued by WRF shows that the prediction accuracy of the input selection-based data-driven models can be improved, especially for long-term predictions (up to 12 hours). Results show that for short lead times (up to 3 hours) the heat flux is the most relevant driver, while for longer lead times a combination of drivers, such as wind and temperature, is selected. Also, such combination varies with the meteorological season.

This approach can thus be adopted to improve the accuracy of rainfall forecasts, although it must be noted that the overall accuracy is still influenced by the underlying numerical model. Indeed, if the numerical weather model does not adequately represent some events (e.g. small-scale convective storms), the selected variables cannot be successfully adopted to yield a rainfall forecast.

Future research will focus on the investigation of the data-driven model performance, which could be enhanced by adopting a hierarchic modelling approach, which aims at firstly predicting the type of storm event, and then running a storm-specific data-driven model. Another research direction stands in evaluating the effect of state variables outside of the catchment, which may be drivers of local rainfall events. Finally, the rainfall forecasts will be included in an operational framework (Galelli et al., 2012), in order to provide long-term inflow predictions to Marina Reservoir and to assess their impact on the barrage operation.

Acknowledgement

This work was carried while the first author was on leave at the National University of Singapore, sponsored by a Politecnico di Milano scholarship. The second author is supported by the SRG ESD 2013 061 Start-up Research Project.

REFERENCES

Baer, F. (2000). Numerical weather prediction. *Advances in Computers 52*, 91–157.

Fong, M. (2012). *The weather and climate of Singapore*. Singapore: Meteorological Service Singapore.

Galelli, S. and A. Castelletti (2013a). Assessing the predictive capability of randomized tree-based ensembles in streamflow modelling. *Hydrology and Earth System Sciences 17*, 2669–2684.

Galelli, S. and A. Castelletti (2013b). Tree-based iterative input variable selection for hydrological modelling. *Water Resources Research 49*(7), 4295–4310.

Galelli, S., A. Goedbloed, D. Schwanenberg, and P. van Overloop (2012). Optimal real-time operation of multi-purpose urban reservoirs: a case study in Singapore. *ASCE Journal of Water Resources Planning and Management*. doi: 10.1061/(ASCE)WR.1943-5452.0000342.

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

Golding, B. (2009). Long lead time flood warnings: Reality or fantasy? *Meteorological Applications 16*, 3–12.

Kristiana, R., J. Antenucci, and J. Imberger (2011). Sustainability assessment of the impact of the Marina Bay development on Singapore: application of the index of sustainable functionality. *International Journal of Environment and Sustainable Development 10*(1), 1–35.

Skamarock, W., J. Klemp, J. Dudhia, D. Gill, D. Barker, M. Duda, X. Huang, W. Wang, and J. Powers (2008). A description of the advanced research WRF version 3. Technical report, National Center for Atmospheric Research, Boulder, CO.

Wang, W., C. Bruyère, M. Duda, J. Dudhia, D. Gill, H. Lin, J. Michalakes, S. Rizvi, X. Zhang, J. Beezley, J. Cohen, and J. Mandel (2011). ARW version 3 modeling system user's guide. Technical report, National Center for Atmospheric Research, Boulder, CO.

Wilson, J., N. Crook, C. Mueller, J. Sun, and M. Dixon (1998). Nowcasting thunderstorms: A status report. *Bullettin of the American Meteorological Society 79*, 2079–2099.

Xie, J. (2006). Dealing with Water Scarcity in Singapore: Institutions, Strategies, and Enforcement. China: Addressing Water Scarcity Background Paper No. 4, The World Bank - Environment and Social Development Department - East Asia and Pacific Region, Washington, D.C.

Zoppou, C. (2001). Review of urban storm water models. *Environmental Modelling & Software 16*(3), 195–231.