Environnement Canada   Environment Canada

Assessing the Skill of Weather Forecasts For the Purpose of Flood Forecasting in Lake Champlain and Richelieu River

Vincent Fortin, Etienne Gaborit and Milena Dimitrijevic

Recherche en Prévision Numérique Environnementale (RPN-E)

Meteorological Research Division

Technical report prepared for the

International Lake Champlain – Richelieu River Technical Working Group

August 31st, 2015

Canada

Information contained in this publication or product may be reproduced, in part or in whole, and by any means, for personal or public non-commercial purposes, without charge or further permission, unless otherwise specified.

You are asked to:

Exercise due diligence in ensuring the accuracy of the materials reproduced;
Indicate both the complete title of the materials reproduced, as well as the author organization; and
Indicate that the reproduction is a copy of an official work that is published by the Government of Canada and that the reproduction has not been produced in affiliation with or with the endorsement of the Government of Canada.

Commercial reproduction and distribution is prohibited except with written permission from the Government of Canada's copyright administrator, Public Works and Government Services of Canada (PWGSC). For more information, please contact PWGSC at 613-996-6886 or at droitdauteur.copyright@tpsgc-pwgsc.gc.ca.

© Her Majesty the Queen in Right of Canada, represented by the Minister of the Environment, 2012

Aussi disponible en français

Table of Contents

Summary

Accurate and reliable wind, precipitation and temperature forecasts can help to forecast floods in Lake Champlain and Richelieu River. Indeed, precipitation falling directly over the lake affects the water level and outflow of Lake Champlain, and leads to increased tributary flow, which further increases water level. Floods can also be triggered by snow melt, which can be forecasted using temperature-index methods. When the water level is already high in Lake Champlain, flooding can be triggered or exacerbated by a wind-induced surge.

An evaluation of Canadian and US wind forecasting systems concludes that it is possible to forecast hourly mean wind speed with a high level of skill at least two days in advance, likely three days. Daily precipitation associated with extra-tropical storms can be forecasted with some skill five days ahead. Degree-day forecasts issued during the months of March and April show better skill than climatology for two weeks. February degree-day forecasts seem to have some skill up to one month into the future.

Acknowledgments

Data presented in this project comes from many sources, and was provided in many different formats. Obtaining the data and making sure that we were interpreting it correctly required major efforts from the research team, but could not have been done without help from many collaborators, in particular members of the International Lake Champlain – Richelieu River Technical Working Group. We wish to acknowledge in particular the contribution of Bill Saunders (Northeast River Forecast Center of NOAA), who provided access to forecasts, precipitation analyses, ancillary data as well as scientific and technical advice. Stephanie Castle (Lake Champlain Basin Program) helped with access to observed wind data, organized teleconferences and took minutes. Ronald Frenette (Environment Canada) provided access to an archive of outputs from the US Short Range Ensemble Forecasts (SREF). Richard Turcotte (Ministère du Développement Durable, de l'Environnement et de la Lutte contre les Changements Climatiques, Gouvernement du Québec) and Jean-François Cantin (Environment Canada) also provided helpful advice. Discussions with colleagues during a workshop organized as part of the 2015 annual conference of the International Association of Great Lakes Research (IAGLR) was also very helpful. We wish to thank the IAGLR local organizing committee and the International Joint Commission (IJC) for helping set up this workshop.

Introduction

With the objective of preparing for the future implementation of a state-of-the art flood forecasting system on Lake Champlain and the Richelieu River, Environment Canada was mandated by the International Lake Champlain – Richelieu River Technical Working Group to evaluate the skill of existing weather forecasts that are available over the watershed of the Richelieu River, including Lake Champlain and its tributaries.

Wind, precipitation and temperature forecasts from existing Canadian and US forecasting systems were evaluated, with the following questions in mind:

- How far ahead in the future can wind, precipitation and temperature be forecasted with skill?
- Does wind forecast skill and bias improve with horizontal resolution?
- How do deterministic and ensemble forecasts of wind and precipitation compare?
- How do US and Canadian wind forecasts compare over Lake Champlain?

The study focused on flood-prone weather conditions:

- sustained North-South winds that cause storm surges
- precipitation events associated with extra-tropical storms
- snowmelt events

Evaluation of wind forecasts

Whereas snowmelt and precipitation events can lead to prolonged flooding by raising the average level of Lake Champlain, wind-induced storm surges are short-lived, as wind only causes an imbalance in water level, with the average water level being almost unchanged. However, sustained strong winds can raise the water level at one end of the lake by tens of centimeters in a matter of hours.

The evaluation of wind forecasts extended from January 2011 through June 2015, and focused on events for which a strong storm surge signal was observed in lake levels.

Wind forecasting systems

Accurate forecasting of surface wind speed requires numerical weather prediction (NWP) systems running at sufficient resolution to resolve the main topographical features and roughness of the surface. In the case of Lake Champlain, which is about 20-km wide at its largest point and bordered on each side by a major mountain range, a resolution of 20-km or higher is deemed necessary. With this constraint in mind, wind speed forecasts from four NWP systems having a horizontal resolution between 2.5-km and 15-km were evaluated against observations:

- the Canadian High-Resolution Deterministic Prediction System (HRDPS), which has a horizontal resolution of 2.5-km and provides forecasts for the next two days (only 18-h prior to October 2014)
- the Canadian Regional Deterministic Prediction System (RDPS), which had a horizontal resolution of 15-km until October 2012 (10-km since then) and provides forecasts for the next two days
- the Canadian Regional Ensemble Prediction System (REPS), which has a horizontal resolution of 15-km and provides forecasts for the next three days
- the U.S. Short Range Ensemble Forecasting system (SREF), which has also a horizontal resolution of 15-km and provides forecasts for the next three days and a half

Whereas the HRDPS and RDPS are deterministic forecasting systems and thus provide a single scenario for wind speed at each grid point, the REPS and SREF both provide twenty scenarios based on stochastic perturbations of initial conditions and model physics, as well as an unperturbed control member. Technical documentation on all Canadian NWP systems can be found online: http://collaboration.cmc.ec.gc.ca/cmc/cmoi/product_guide/.

Only the RDPS model outputs were available over the whole period of interest (January 2011 to June 2015). Forecasts from the HRDPS were available from April 2012. REPS forecasts were only available at 15-km starting in 2014, and SREF forecasts were only available for 2015. Hence, in order to answer the science questions on wind forecast skill, the RDPS was considered as the reference model:

- the RDPS forecasts were used to first assess how forecast skill changes with lead time
- the HRDPS forecasts were then compared to the RDPS forecasts in order to assess the impact of horizontal resolution (2.5-km vs 10-km and 15-km)
- the SREF and REPS models were compared to the RDPS in order to assess the added value of ensemble forecasts
- the SREF and REPS forecasts were finally compared in order to see how forecast skill compares in US and Canada

It must be stressed that all Canadian NWP systems rely on different configurations of the same atmospheric model, GEM (Côté et. al 1998; Girard et al., 2014). Although there are differences other than horizontal resolution between these configurations, the other changes are essentially made to adapt the model physical parameterizations to the horizontal resolution. For this reason, a comparison of Canadian NWP systems provides a meaningful way of assessing the impact of horizontal resolution on model bias and skill. Note also that the HRDPS forecasts are obtained by dynamical downscaling of RDPS forecasts, and are hence highly correlated, which is an advantage here in order to isolate the impact of horizontal resolution when comparing forecasts.

Wind observations

Wind observations from four weather stations were initially considered: Burton Island, Colchester Reef, Burlington Airport, and Diamond Island (Figure 1). The three yellow dots represent wind gauges for which data comes from the University of Vermont. The red one comes from the National Oceanic and Atmospheric Administration (NOAA) National Climate Data Center (NCDC). More information on these stations is provided in Table 1.

In order to assess the quality of the wind forecasts over Lake Champlain, it was finally chosen to use, as observations, the wind data from the Vermont Monitoring Cooperative (VMC) network of the University of Vermont. These data correspond to 15 min. averages of wind speed and direction and were measured between 4 and 19-m above ground depending on the site considered (see Table 1). The Burlington station, located over land, was not used in the end for the evaluation as its observed data were judged less relevant than the three Vermont Monitoring Cooperative gauges in terms of wind conditions prevailing over Lake Champlain.

Furthermore, a detailed assessment of the wind speed observations revealed that wind speed observations remained very low and almost constant over relatively long periods of time at the Burton Island and Diamond Island stations. The team responsible for the wind measurements was contacted, and they confirmed that trees located directly north of the stations could protect them from north winds, preventing data to be accurately measured for such events. For the events considered, this seemed to occur more frequently for the Burton Island station.

Colchester Reef data, situated near the middle of the lake, hence provides the most interesting dataset for model verification, and will therefore be used as the reference of choice for model forecast evaluation.

Table 1: Wind station data characteristics

Gauge	LAT	LON	Elevation (m)	Data source	original units	time interval
Colchester Reef	44.5542	-73.3285	7	http://www.uvm.edu/vmc/	m/s	15 min.
Diamond Island	44.2367	-73.3329	19		m/s	15 min.
Burton Island	44.7666	-73.2134	4		m/s	15 min.
Burlington international Airport (BTV) 14742	44.468	-73.149	?	https://www.ncdc.noaa.gov/data-access/land-based-station-data/land-based-datasets/quality-controlled-local-climatological-data-qclcd	MPH	5 min. - hourly

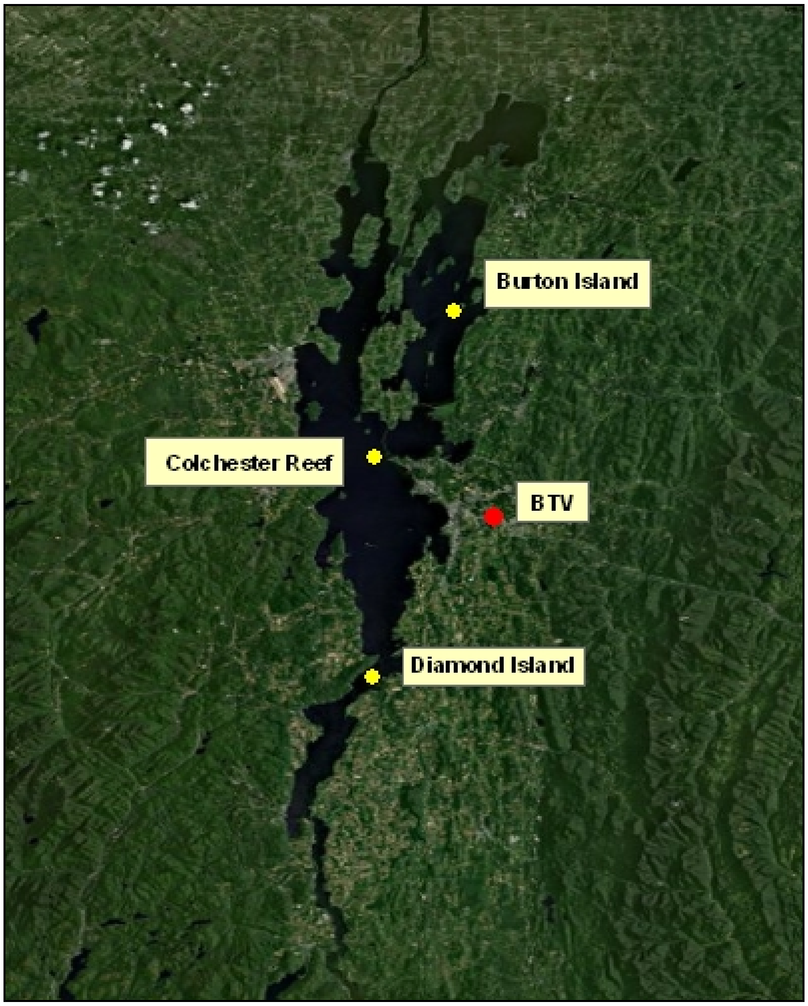


Figure 1: location of the wind gauges

Water level observations

Lake level data from six stations were obtained (Figure 2) from Water Survey of Canada and USGS websites.

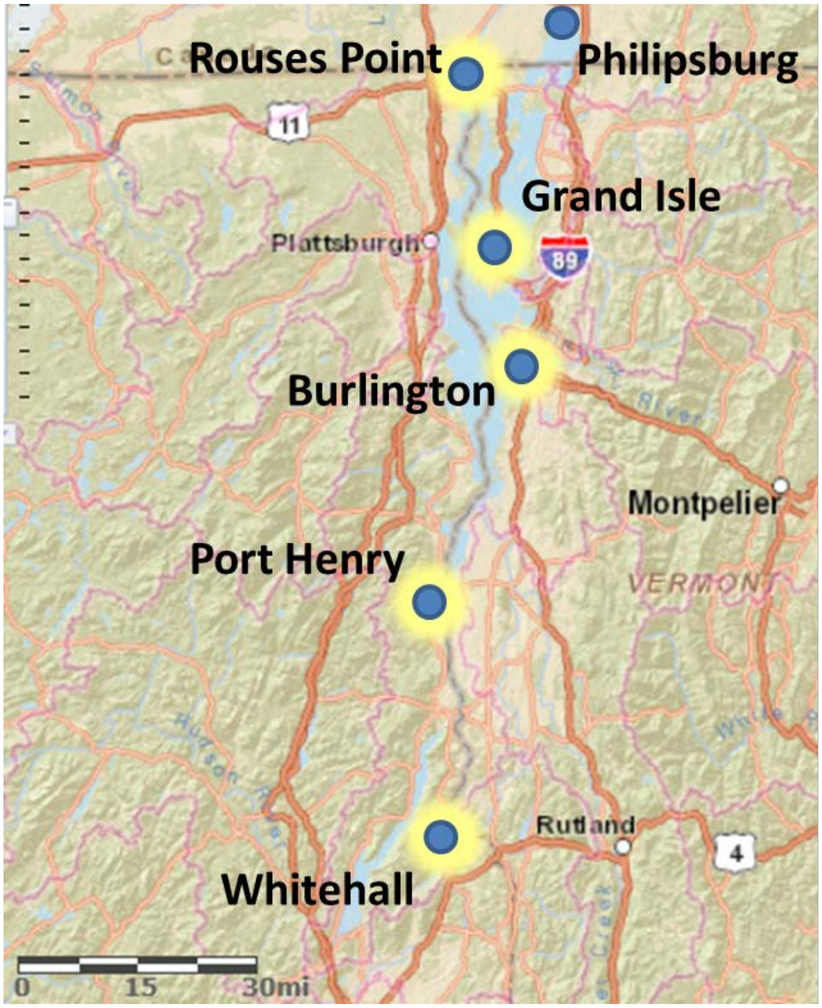


Figure 2: location of the water level gauges

Data was available for four of them for the duration of the study:

- Philipsburg, located in the north-east bay north of the U.S./Canadian border
- Rouses point, located near the outlet of the lake
- Burlington, located near the center of the lake
- Whitehall, located near the south end of the lake, in a very narrow section

Two more stations became available during the spring of 2015:

- Port Henry, located near the south of the lake
- Grand Isle, located near the south end of the Inland Sea

The Philipsburg and Grand Isle stations are located in the Inland Sea, an area of Lake Champlain which is only connected to the main part of the lake through narrow causeways, and hence can experience different water level conditions.

Identification of storm surge events

Wind forecast skill can vary tremendously, depending on the weather conditions. For example, large-scale events will often be better forecasted than wind caused by localized storms. In order to evaluate skill for events that matter for flood forecasting, storm surge events were identified based on a visual inspection of water level data. Figure 3 illustrates how water level can vary in time and space during a typical storm surge event on Lake Champlain. It shows 15-min water level changes at five stations between 2015-05-30 00 UTC and 2015-06-01 00 UTC. Burlington station, which is near the middle of the lake, typically does not experience any large changes in water level due to wind, and is not shown. Table 2 presents the Pearson correlation coefficient between each pair of stations during this event. For this event, Port Henry and Philipsburg display the strongest negative correlation (-0.90), whereas Whitehall and and Port Henry display the strongest positive correlation (0.71).

Figure 3: Water level fluctuations during a storm surge event on Lake Champlain

Table 2: Pearson correlation coefficient of 15-min observations for each pair of water level stations between 2015-05-30 00 UTC and 2015-06-01 00 UTC

	Whitehall	Port Henry	Rouses Point	Grand Isle	Philipsburg
Whitehall	1.00	**0.71**	-0.46	0.39	-0.80
Port Henry	0.71	1.00	-0.81	0.65	**-0.90**
Rouses Point	-0.46	-0.81	1.00	-0.81	0.70
Grand Isle	0.39	0.65	-0.81	1.00	-0.54
Philipsburg	-0.80	-0.90	0.70	-0.54	1.00

Note that the water level shown on Figure 3 is the difference from the start of the event. Over two days, four out of five gauges come back within 1 cm of their initial level. Hence we can say with some confidence that changes in water level were mostly due to wind effect at these four locations. Whitehall has markedly different behavior, possibly reacts to things other than wind. Indeed, the correlation between the Port Henry gauge and the other gauges is always higher than the correlation between the Whitehall gauge and the other gauges. Given the fact that the Whitehall gauge is located quite far from the main body of Lake Champlain, in a very narrow strait, it therefore seems that this storm surge event, and possibly others, would be best characterized quantitatively by looking at differences between the Port Henry gauge and either the Rouses Point or the Philipsburg gauge. Unfortunately, the Port Henry gauge only became available in 2015. This event also shows some evidence that a different wind set-up event is occurring in the inland sea. Indeed, the Grand Isle gauge, located in the northern half of the lake but at the southern end of the inland sea, is positively correlated with the southern gauges at Porth Henry and Whitehall.

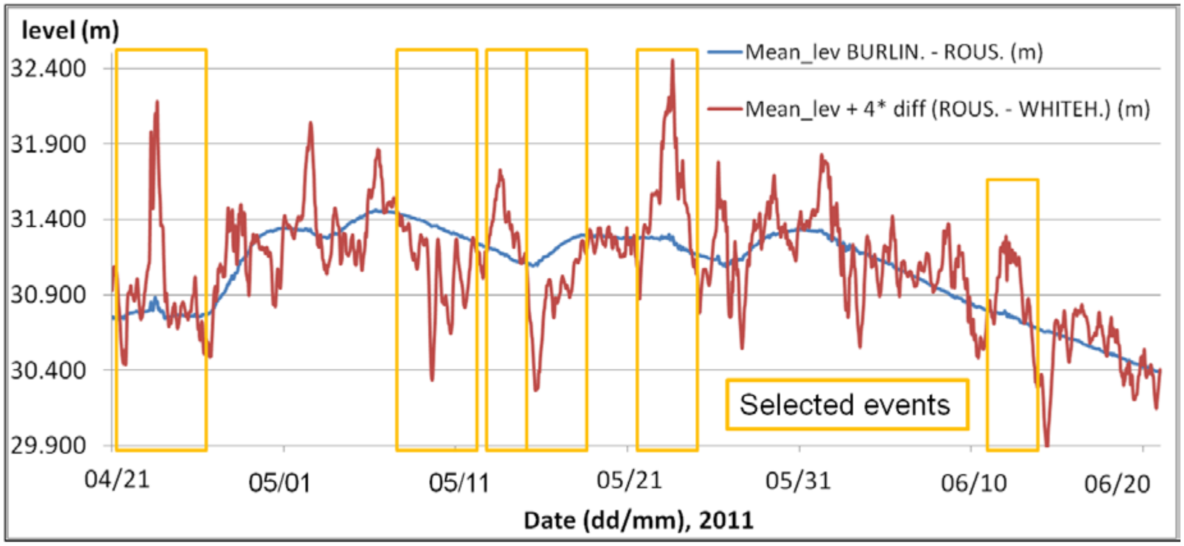


Figure 4: plot of the lake Champlain mean level (m) and mean level plus 4 times the difference between its extremities (Rouses Point and Whitehall), to identify strong wind events having an effect on lake levels. The figure displays only a sample of the full 2011-2014 period used to select the events. The 6 highlighted events were selected for the wind forecast evaluation.

Nonetheless, a visual inspection of differences between water level over the period of 2011-2014 at its extremities (Rouses Point and Whitehall) did enable us to identify storm surge events, as illustrated by Figure 4. Over the course of these four years, thirty-six events were identified in this manner, all corresponding to a period of sustained winds in the North-South direction. A list of these events is presented in Annex 1.

Evaluation methodology

Because the lake is narrow and elongated in the North-South direction, the focus was put on the evaluation of the North-South component of the wind, which is expected to have the largest impact on wind set-up. Wind blowing from the South towards the North is defined here to be positive, and wind from the North to be negative. A positive North-South wind hence means that water is being pushed towards the outlet of the lake, and contributes to increased flow in the river.

The forecast quality of each model was assessed as a function of the forecast lead-time for the model pixel closest to each station. Different scores were used:

- the mean error (forecasted wind minus observed) or bias
- the Pearson correlation coefficient
- the mean absolute error (MAE)
- the root mean square of errors (RMSE)

In the case of ensemble forecasting systems, these scores were computed both for the ensemble mean and selected quantiles.

The Continuous Ranked Probabilistic Score (CRPS, see Matheson and Wrinkler 1976) was also computed for ensemble forecasting systems. It is a probabilistic score which considers all ensemble forecast members at the same time. It can be directly compared to the MAE, its deterministic equivalent. When the CRPS is smaller than the MAE of the ensemble mean , this indicates that the ensemble members provide a potentially useful description of the forecast uncertainty. Moreover, the CRPS score can be decomposed, following Hersbach (2000), into a potential CRPS (CRPS_pot) and a reliability component (CRPS_reli). CRPS_pot corresponds to the CRPS that could ideally be obtained if the ensemble forecasts had a perfect reliability, i.e. if they correctly described the forecast uncertainty, including the bias of the ensemble mean. CRPS_reli is the part of the total CRPS which can be attributed to the unreliability of the ensemble product. In terms of interpretation, if the CRPS is high and close to CRPS_reli, this means that the product is biased and that there is room to improve it.

Evaluation of RDPS wind forecasts

As mentioned earlier, data from the Colchester Reef station was used as the main reference for model forecast evaluation. 2162 hourly forecasts were evaluated over the course of the 36 selected events (about 90 days in total). Hourly forecasts were compared against 15-min averages of observed winds, (which is how the data is reported by VMC). Indeed, even if forecast outputs are available at a hourly time step, the internal model time step is smaller and the forecasted wind corresponds to an instantaneous wind (albeit without any wind gust component).

Surface winds from the RDPS are available at two levels: 40-m and 10-m. 40-m corresponds to the lowest vertical level of the model, whereas 10-m is a model diagnostic obtained by extrapolating wind speed closer to the surface based on Monin-Obukhov similarity theory. It is interesting to see how both forecasts compare to observed winds for different lead times (Table 3).

Table 3: performances of the RDPS wind forecasts (North/South component) for the 36 selected events vs Colchester Reef. Errors are in km/h.

		RDPS 40m			RDPS 10m		
horizon	lead-time (h)	Corelation coeff	Mean absolute error	Mean error	Corelation coeff	Mean absolute error	Mean error
1	1-6	0.94	8.36	-5.07	0.94	12.48	-10.81
2	7-12	0.94	8.70	-5.12	0.93	12.69	-10.82
3	13-18	0.93	8.92	-5.29	0.93	12.87	-11.01
4	19-24	0.92	9.10	-4.91	0.92	12.89	-10.71
5	25-30	0.92	9.10	-4.52	0.92	12.73	-10.43
6	31-36	0.91	9.46	-4.74	0.91	12.99	-10.57
7	37-42	0.90	9.89	-4.86	0.90	13.26	-10.70
8	43-48	0.90	9.99	-4.76	0.90	13.33	-10.61

It can be seen that 10-m and 40-m winds show the same correlation with observed winds, which is expected since the 10-m wind is a diagnostic wind. However, the 10-m wind forecasts show a stronger downward bias (or mean error) than 40-m wind forecasts. It is normal for the 10-m wind forecast to be of lower intensity than the 40-m wind forecast. At Colchester Reef, they are on average lower by about 6 km/h. In this case, because the 40-m wind is already biased downwards (with a mean error around -5 km/h), the extrapolation to 10-m degrades the forecast bias to around -11 km/h. However, the correlation remains quite high (0.90 or higher two days ahead).

It therefore seems that forecasts of comparable skill are obtained from the RDPS at very short range and at longer range (two days ahead). This is confirmed by a visual inspection of the forecasted wind speed at Colchester Reef for two important events of 2011: the historical spring freshet and Hurricane Irene (Figure 5). It can be seen that 40-m forecasts captured these events quite well, both for the first forecast horizon (1 to 6 hours) and for the last forecast horizon (43 to 48 hours). In both cases, an underestimation of peak wind speed is observed, but the timing and magnitude of the forecast both seem quite acceptable.

It should however be stressed that Figure 5 compares 7-m observations with 40-m forecasts. The bias of the forecast is here artificially reduced by considering a wind forecast higher up in the atmosphere, and is reason for concern if the objective is to couple the NWP system to a hydrodynamic model.

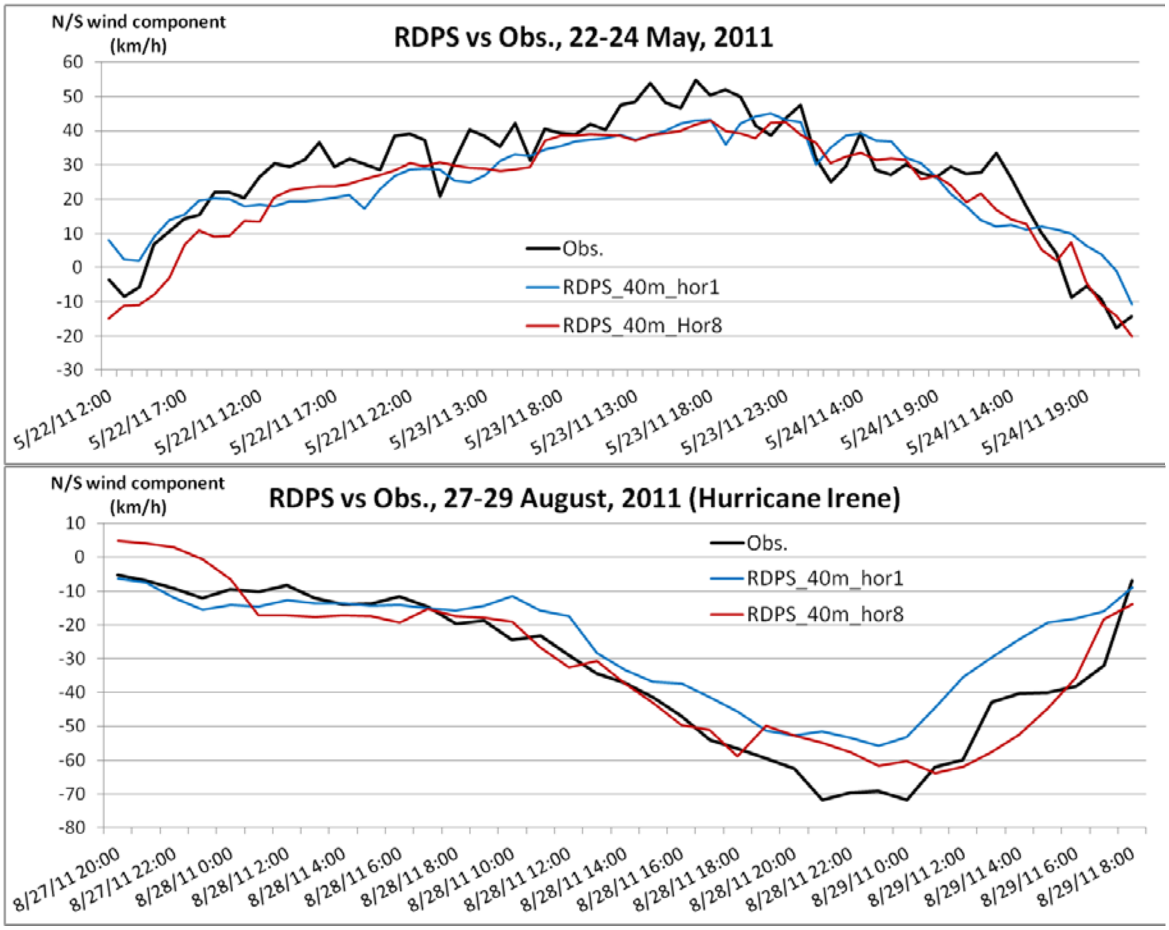


Figure 5: Comparison between the first (0-6h) and last (42-48h) horizons of the RDPS 40-m outputs. The figure shows the N/S component of wind speeds (km/h) at Colchester reef station, for two main events of the 2011 period (top and bottom graphs).

One possible explanation for the model bias is the horizontal resolution of the model, which changed from 15-km to 10-km over the course of the evaluation period (in October 2012). Other model configurations occurred in November 2011 and February 2013. To see if these changes (including horizontal resolution) had a positive impact of model performance, the dataset was split, and scores were computed separately for forecasts issued before November 2011 (1054 forecast hours considered) and after February 2013 (742 forecast hours considered).

Results of this comparison are presented in Table 4. It can be seen that although a modest improvement in Mean Absolute Error is observed, the Mean Error (or bias) was not reduced. In fact, it increased. The resolution change from 15-km to 10-km clearly did not eliminate the model bias. However, for a lake narrow as Lake Champlain, it can be argued that a model resolution of 10-km is still insufficient to resolve the topography, roughness and channeling effects. A comparison with the 2.5-km HRDPS system is therefore required to better address this issue.

Table 4: Influence of RDPS main upgrades on performances of 40-m wind forecasts, for the Colchester reef station.

	Mean_Error		Mean_abs_Error	
	before Nov_2011	After Feb_2013	before Nov_2011	After Feb_2013
hor 1	-5.37	-6.26	7.90	7.98
hor 3	-6.00	-6.31	8.58	8.01
hor 6	-5.73	-5.70	9.05	8.14
hor 8	-5.82	-6.62	9.82	8.45

Evaluation of HRDPS wind forecasts

The HRDPS data being only available from April 2012, the comparison between the RDPS and HRDPS forecasts are therefore done on a subset of the 36 events, containing a total of 1003 forecast-observation pairs. At the time, HRDPS forecasts were only available for lead times of 13 to 36 hours. The objective of this comparison being to assess the impact of horizontal resolution, forecasts for all lead times were combined, in order to more accurately measure the difference in model bias when increasing the horizontal resolution of the model. Table 5 summarizes the results. Although no increase in the correlation coefficient was observed, an important reduction of the mean error (or bias) was observed, along with reductions in MAE and RMSE. The fact that the correlation coefficient was unchanged suggests that the reductions in MAE and RMSE were essentially due to a reduction in bias.

Table 5: Comparison between RDPS and HRDPS 10-m outputs for the Colchester reef station, over the April 2012-October 2014 subset of events and lead-times between 13 and 36 hours.

	HRDPS	RDPS
Corelation coeff.	0.93	0.93
Mean error	-4.43	-11.65
Mean absolute error (MAE)	8.25	13.21
RMSE	11.49	15.26

This comparison suggests that higher resolution is definitely useful for improving forecast bias and skill, but that statistical bias correction of lower-resolution forecast could be also considered as a viable option to obtain forecasts of similar skill. Note however that the comparison is performed here against a weather station located near the center of the lake, where the lower resolution model has the best chance of simulating the wind flow well. Closer to the coast or near islands, it is possible that the HRDPS would have more of an edge over its lower resolution counterpart.

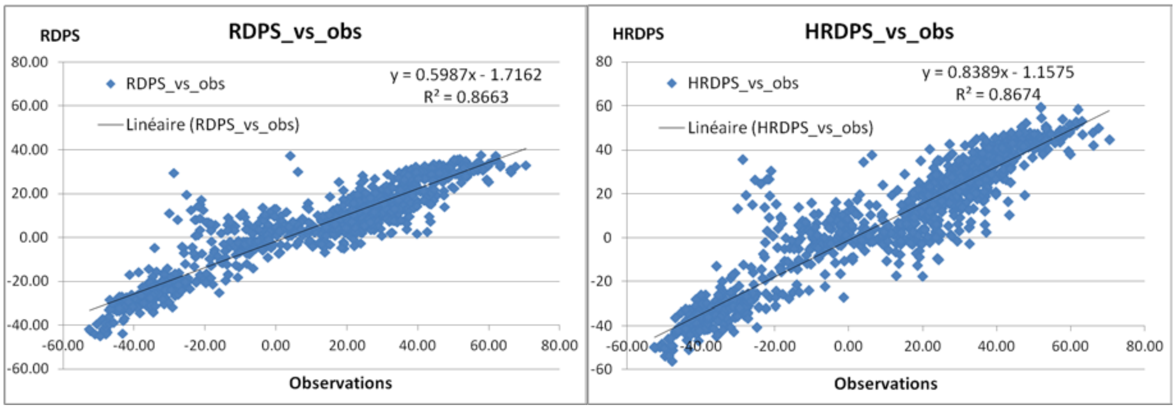


Figure 6: RDPS (left panel) and HRDPS (right panel) 10-m N/S component of wind speeds (km/h), versus observed N/S component of wind at Colchester reef station, over the 2012-2014 period (see text).

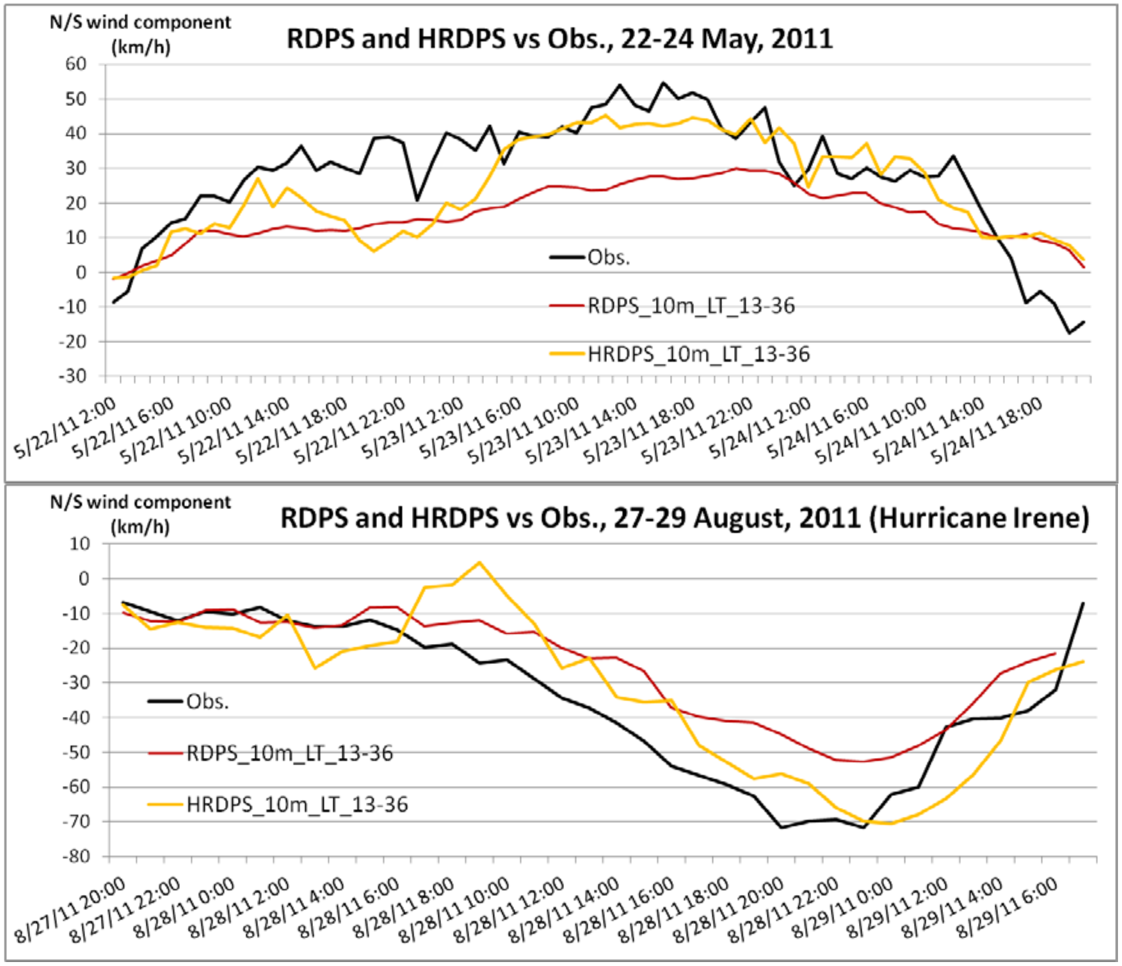


Figure 7: Comparison between 10-m North/South wind forecasts from the RDPS and GEM-LAM East (replaced by the HRDPS) for two main events of the 2011 period (top and bottom graphs).

The improvement in bias is obvious when considering scatter plots of these forecasts (Figure 6). Although the correlation remains almost the same when going from 10-15 km down to 2.5-km, the magnitude of forecasted winds is increased, especially for winds blowing from the South which can exacerbate flooding downstream: with the RDPS, no wind speed higher than 40 km/h was ever forecasted over this period, whereas the HRDPS forecasts go up to 60 km/h.

Although forecasts from the HRDPS were not available in 2011, an experimental system was already in place which had the same horizontal resolution of 2.5-km (GEM-LAM east), and which gives a good idea of what the current HRDPS system could provide in terms of forecast skill. Figure 7 shows the 10-m forecast issued by the RDPS and GEM-LAM east for the two main flooding events of 2011. Although the GEM-LAM east bias is clearly smaller than the RDPS bias for these two events, correlation with observed wind does not appear to be improved. Note that 10-m wind forecasts are shown on this figure, compared to 40-m wind forecasts for Figure 5.

It can be noted that wind forecasts on Figure 5 seem to lag observed wind by 2-3 hours, especially later in the forecast. Since forecasts with lead times of 13-36 hours are compared, the lag is mostly apparent in the day-2 forecast. Such timing errors can be expected, due in this case to the fact that the extra-tropical storm Irene moved slightly faster than forecasted.

Evaluation of SREF and REPS wind forecasts

With SREF forecasts being available only from the beginning of 2015 and up to June 2015, it was decided to identify a new set of events that would take advantage of new water level stations that came online in 2015 (Port Henry and Grand Isle), so that the exact same set of events used for the wind evaluation performed in this section could also be used to test a future hydrodynamic model of Lake Champlain with all the possible water gauges available for the area. In this sense, it is not only the SREF forecasts, but also the water level gauges' data availability, which together determined the extent of the period used in this new evaluation: April through June 2015. The events were selected as before, namely by looking at strong wind events mainly oriented in the N/S direction and verifying that these events translated into lake level differences. Despite being short, this 3 month period luckily involved several strong wind events, both with Northerly and Southerly winds. Annex 1 presents a synthesis of the events selected for this new wind evaluation.

However, because of limitations in the SREF data availability, and because data had to also be available two days prior to an event start date (in order to compute performances up to a 48-h lead-time), the evaluation was conducted only for the days presented in Table 6, which together contain at least a part of all selected events shown in Annex 1.

Table 6: days for which the evaluation was made possible due to limitations in the SREF data availability.

2015		
April	May	June
17-20	12-13	4-9
	18-21	
	29-31	

Moreover, as the SREF forecasts were available only at a 3-h time-step and only for the forecasts issued at 9 UTC, the number of observation/forecast pairs was further reduced and, in the end, equal to only 124 pairs (15.5 days), which is very short compared to the data used previously to evaluate the 10-km and 15-km deterministic forecasts (between 1000 and 2200 time-steps). Since the idea here is mainly to compare the products together rather than having a precise idea of the magnitude of the errors associated with the wind forecasts, this was judged sufficient.

Tables 7-10 present the results of this model inter-comparison study, for the 124 forecast/observation pairs of the selected wind events, for both the Colchester Reef and Diamond Island stations. Burton Island was not used, because the 2015 data at this station presented a significant fraction of a priori erroneous values.

It is interesting to first compare the RDPS and the HRDPS over this second period. Recall that over the period of April 2012 – December 2014, the HRDPS and RDPS showed a similar correlation to Colchester Reef observations, with the HRDPS providing the less biased forecasts. During these three months of 2015, the HRDPS was less biased than the RDPS at both Colchester Reef and Diamond Island stations, but this time it provides a slightly improved correlation for Colchester Reef, and a slightly deteriorated correlation for Diamond Island. Given the small comparison period and the sensitivity of the Pearson correlation coefficient to outliers in small samples, results can be considered consistent with what was observed in the first comparison period.

It is then striking to observe that the much lower resolution SREF model displays a bias that is similar to the HRDPS bias, and a bias that is better than that of the REPS and RDPS bias for both lead times and both stations. This is true for the control member of the SREF and for the ensemble mean.

The MAE and RMSE are generally better for the HRDPS than for the SREF at Colchester Reef, but the opposite is generally true for Diamond Island. And at both stations, SREF forecasts are better correlated with observations than all Canadian NWP systems. This suggests that for wind forecasting on Lake Champlain, the US SREF system currently provides the best forecasts, with a bias comparable to that of the Canadian HRDPS system, and a better accuracy.

Table 7: Results of the evaluation of day one forecasts using the events of 2015, for the N/S component of the wind at 10-m (km/h), against the Colchester Reef station. Abs: absolute value. Hor: Horizon. CTL: Control. MED: median. Max: maximum. For the ensemble products (SREF and REPS), the evaluation was performed for different quartiles, namely the median (50%), third (75%) and last quartiles (maximum value of the members). CRPS_TOT: Continuous Ranked Probabilistic Score. CRPS_POT: potential CRPS. Values for the SREF+REPS hybrid product were obtained by merging both products' members, in which case the Control member is the mean of both products' control members.

		Hor1 (0-24h)						
		CTL or MEAN(CTL)	MEAN	50 (MED)	75	max	CRPS_TOT	CRPS_POT
MEAN ABS. ERROR (MAE)	SREF	9.95	10.36	10.47	8.92	6.60	8.65	1.99
	REPS	13.10	13.00	12.99	11.82	9.85	11.81	0.76
	SREF+REPS	11.44	11.52	11.80	9.49	6.50	9.60	2.10
	RDPS	12.70						
	HRDPS	9.74						
RMSE	SREF	11.57	12.09	12.24	10.60	8.38		
	REPS	15.49	15.4	15.38	14.26	12.27		
	SREF+REPS	13.27	13.47	13.76	11.33	8.36		
	RDPS	14.24						
	HRDPS	11.66						
COREL	SREF	0.76	0.78	0.78	0.78	0.73		
	REPS	0.65	0.65	0.66	0.66	0.64		
	SREF+REPS	0.75	0.75	0.76	0.79	0.73		
	RDPS	0.67						
	HRDPS	0.70						
MEAN ERROR (abs(sim)-abs(obs))	SREF	-8.46	-9.22	-9.45	-7.29	-1.26		
	REPS	-12.49	-12.34	-12.35	-10.93	-8.00		
	SREF+REPS	-10.47	-10.64	-11.02	-8.26	-0.99		
	RDPS	-11.10						
	HRDPS	-7.76						

Table 8: Same as Table 7, but for day two of the forecast against the Colchester Reef station

		Hor2 (24-48h)						
		CTL or MEAN(CTL)	MEAN	50 (MED)	75	max	CRPS_TOT	CRPS_POT
MEAN ABS. ERROR (MAE)	SREF	10.75	11.24	11.36	9.69	7.33	9.35	2.42
	REPS	11.87	12.02	12.01	10.87	8.73	10.67	1.21
	SREF+REPS	11.07	11.53	11.76	9.89	6.91	9.69	2.57
	RDPS	12.80						
	HRDPS	9.55						
RMSE	SREF	12.81	13.46	13.61	11.72	9.26		
	REPS	14.28	14.44	14.42	13.32	10.82		
	SREF+REPS	13.32	13.83	14.07	12.19	8.70		
	RDPS	14.53						
	HRDPS	11.82						
COREL	SREF	0.70	0.68	0.69	0.67	0.65		
	REPS	0.62	0.65	0.65	0.64	0.66		
	SREF+REPS	0.70	0.68	0.68	0.67	0.69		
	RDPS	0.61						
	HRDPS	0.67						
MEAN ERROR (abs(sim)-abs(obs))	SREF	-9.45	-10.09	-10.37	-7.60	-0.91		
	REPS	-10.70	-11.12	-11.07	-9.52	-5.82		
	SREF+REPS	-10.08	-10.56	-10.88	-8.26	-0.16		
	RDPS	-11.00						
	HRDPS	-7.35						

Table 9: Same as Table 7 but for the Diamond island station for day one forecasts

		Hor1 (0-24h)						
		CTL or MEAN(CTL)	MEAN	50 (MED)	75	max	CRPS_TOT	CRPS_POT
MEAN ABS. ERROR (MAE)	SREF	10.28	10.99	11.18	9.49	6.79	9.23	2.01
	REPS	14.49	14.37	14.36	13.38	11.77	13.40	0.49
	SREF+REPS	12.30	12.47	13.00	10.44	6.77	10.58	2.04
	RDPS	13.50						
	HRDPS	12.07						
RMSE	SREF	12.13	12.88	13.12	11.28	8.62		
	REPS	17.01	16.86	16.83	15.88	14.26		
	SREF+REPS	14.38	14.59	15.20	12.46	8.59		
	RDPS	15.86						
	HRDPS	14.54						
COREL	SREF	0.74	0.75	0.74	0.76	0.72		
	REPS	0.62	0.64	0.65	0.65	0.62		
	SREF+REPS	0.72	0.73	0.72	0.75	0.72		
	RDPS	0.63						
	HRDPS	0.57						
MEAN ERROR (abs(sim)-abs(obs))	SREF	-8.94	-9.89	-10.16	-7.93	-1.70		
	REPS	-14.00	-13.87	-13.85	-12.71	-10.52		
	SREF+REPS	-11.47	-11.70	-12.34	-9.31	-1.61		
	RDPS	-12.74						
	HRDPS	-10.60						

Table 10: Same as Table 7 but for the Diamond island station for day two forecasts

		Hor2 (24-48h)						
		CTL or MEAN(CTL)	MEAN	50 (MED)	75	max	CRPS_TOT	CRPS_POT
MEAN ABS. ERROR (MAE)	SREF	10.97	11.73	11.89	10.07	7.83	9.83	2.44
	REPS	13.79	13.81	13.79	12.84	11.09	12.75	0.68
	SREF+REPS	12.20	12.59	13.15	11.05	7.74	10.82	2.50
	RDPS	13.50						
	HRDPS	11.69						
RMSE	SREF	13.13	14.01	14.24	12.18	9.58		
	REPS	16.24	16.27	16.23	15.35	13.61		
	SREF+REPS	14.46	14.97	15.62	13.30	9.51		
	RDPS	16.02						
	HRDPS	14.15						
COREL	SREF	0.67	0.65	0.65	0.66	0.63		
	REPS	0.61	0.64	0.64	0.63	0.59		
	SREF+REPS	0.68	0.66	0.64	0.66	0.63		
	RDPS	0.60						
	HRDPS	0.53						
MEAN ERROR (abs(sim)-abs(obs))	SREF	-9.62	-10.57	-10.83	-8.13	-1.41		
	REPS	-13.01	-13.13	-13.09	-11.91	-9.38		
	SREF+REPS	-11.31	-11.74	-12.42	-9.63	-1.29		
	RDPS	-12.74						
	HRDPS	-9.74						

By comparing CRPS_TOT (Total CRPS) to the MAE of the ensemble mean for the SREF and for the REPS, it is possible to confirm that ensemble members do contribute to a useful description of the forecast error. Indeed, CRPS_TOT is always smaller than MAE for the same model. Furthermore, the total CRPS of the SREF is always smaller than the MAE of the HRDPS, for both stations and both lead times, confirming that ensemble forecasts, despite being issued at lower resolution than deterministic products, can provide useful information on wind speed for flood forecasting purposes.

A multi-model ensemble forecast was attempted by combining SREF and REPS forecasts. This leads to an ensemble forecast having a skill in between SREF and REPS performance, and unfortunately not to a better ensemble forecast. This was expected because the sign of the bias is the same for the SREF and REPS. Multi-model ensemble forecasts are particularly useful when models have different biases, as the combined ensemble has typically less bias and more spread.

For both the SREF and REPS, the potential CRPS (CRPS_POT) is much smaller than the total CRPS (CRPS_TOT), which indicates that the reliability component of the CRPS is high. This corresponds both to a bias in the mean and in the spread of the ensemble, something that can potentially be improved through statistical post-processing of the ensemble forecasts. It is also interesting to note that the REPS potential CRPS is actually lower than the SREF potential CRPS, meaning that after calibration, the Canadian REPS could potentially provide wind forecasts of skill similar or better than that of the US SREF system.

Different approaches to bias correction can be attempted. Since all of the wind forecasts considered in this study actually have a negative bias, a possible way of calibrating an ensemble forecasts consists of using as a deterministic forecast a high quantile of the ensemble distribution. This idea was tested by computing scores for the median, 75% quantile, as well as for the maximum of the ensemble (Tables 7-10). Although it is not recommended to use the maximum of an ensemble as a deterministic forecast due to the lack of robustness of this prediction approach, it does in this case seem to considerably reduce the bias of the forecast without unduly affecting its correlation with observed wind speed. But the bias stays negative even in this case.

An alternative but equally simple bias-correction technique was therefore considered for the REPS: each forecast was multiplied (using the same coefficient for all members) by the ratio of the mean observed wind speeds over the mean forecasted wind speeds for the control member of the REPS (using the absolute values of the wind N/S component at a station location). For the Colchester Reef and Diamond Island stations (and for the events studied here), the ratio was equal to 2 and 1.82, respectively. All REPS members were multiplied by these factors, and the post-processed product resulting from this very simple bias-correction trial are presented under the name (REPS_PRD) in Tables 11 and 12. It can be seen that the performances of the resulting product are better than any of the other products, including the HRDPS. This means that it should be possible to considerably improve the forecasts of the different products by using more-sophisticated bias correction techniques.

Table 11: Performance of Canadian NWP systems assessed for Colchester Reef. The total number of forecast/observation pairs is 362 in this case. See legend of Table 7 for the meaning of the different deterministic interpretations of ensemble members. CTL: Control. CRPS_TOT: Continuous Ranked Probabilistic Score. CRPS_POT: Potential CRPS (see section 5). Results shown here are for a forecast lead-time comprised between 0 and 12h. REPS_PRD: bias-corrected REPS forecasts.

		CTL	MEAN	50 (MED)	75	max	CRPS_TOT	CRPS_POT	CRPS_RELI
MEAN ABS. ERROR (MAE)	REPS	13.26	13.19	13.19	12.21	10.41	12.19	0.70	11.49
	REPS_PRD	8.05	7.68	7.82	7.80	9.71	5.77	4.16	1.61
	RDPS	11.89							
	HRDPS	9.10							
RMSE	REPS	15.43	15.36	15.36	14.36	12.44			
	REPS_PRD	10.10	9.70	9.86	9.92	12.28			
	RDPS	13.80							
	HRDPS	11.01							
COREL	REPS	0.70	0.71	0.71	0.71	0.70			
	REPS_PRD	0.70	0.71	0.71	0.71	0.70			
	RDPS	0.76							
	HRDPS	0.75							
MEAN ERROR (abs(sim)-abs(obs))	REPS	-12.63	-12.62	-12.61	-11.38	-8.73			
	REPS_PRD	-1.03	-1.00	-0.97	1.48	6.78			
	RDPS	-11.17							
	HRDPS	-7.40							

Table 12: Same as Table 11, but for Diamond Island station.

		CTL	MEAN	50 (MED)	75	max	CRPS_TOT	CRPS_POT	CRPS_RELI
MEAN ABS. ERROR (MAE)	REPS	9.31	9.26	9.27	8.75	8.06	8.51	1.23	7.28
	REPS_PR	6.44	6.36	6.41	6.41	7.68	4.90	3.17	1.73
	RDPS	8.14							
	HRDPS	7.10							
RMSE	REPS	11.58	11.58	11.57	10.92	9.89			
	REPS_PR	8.18	8.11	8.13	8.29	10.10			
	RDPS	10.20							
	HRDPS	8.70							
COREL	REPS	0.72	0.72	0.71	0.72	0.69			
	REPS_PR	0.72	0.72	0.71	0.72	0.69			
	RDPS	0.75							
	HRDPS	0.74							
MEAN ERROR (abs(sim)-abs(obs))	REPS	-7.86	-7.79	-7.78	-6.83	-4.83			
	REPS_PR	-0.07	0.07	0.08	1.81	5.45			
	RDPS	-6.28							
	HRDPS	-3.90							

Because so few events were considered for the evaluation of the ensemble forecasting systems, it is feasible to look at time series of forecasts for all events in a single plot. Figures 8-9 compare SREF and REPS forecasts (75% quantile), as well as the multi-model ensemble forecast (again the 75% quantile) to observed North-South wind speed at Colchester Reef and Diamond Islands for the events considered in the Spring of 2015. All forecasts show a comparable underestimation of maximum wind speeds, but correctly capture the general tendency. Note that 1-day forecasts are shown on Figure 8, and 2-day forecasts on Figure 9. A comparison with the deterministic NWP systems (RDPS and HRDPS) is shown on Figures 10-11. For both stations, the HRDPS seems to do better for the highest winds, but all systems seem to provide a similar signal, which once again suggests that 15-km NWP systems can provide useful information on wind for flood forecasting purposes.

Summary of the wind forecast evaluation

After evaluating wind forecasts from four NWP systems against two weather stations located on Lake Champlain, it does appear that hourly wind speed can be forecasted with skill at least two days ahead, but that bias forecasts are obtained. Although bias is reduced when horizontal resolution is increased, all forecasting systems considered underestimate wind speed. Short-range ensemble forecasting systems perform well, especially the US Short Range Ensemble Forecasting System (SREF), with performances similar to the High-Resolution Ensemble Forecasting System (HRDPS) from Canada. However, after implementation of a simple bias correction procedure, the more biased Canadian Regional Ensemble Prediction System (REPS) outperforms both systems. Hence, ensemble forecasting of wind for flood forecasting purposes is feasible with existing operational systems, but statistical bias correction procedures should be considered to improve the forecast.

In order to compare ensemble and deterministic products, only forecasts with lead times of 1-h to 48-h were considered. However, ensemble forecasts at the same horizontal resolution are available for at least 72-h. Given the slow drop in correlation with lead time, ensemble forecasts with lead times of 72-h are expected to have useful skill as well.

The weather forecasts from operational ensemble and deterministic ensemble prediction systems are available at no cost from EC and NOAA. However, the computing costs required to run hydrodynamic models from all of these forecasts can be quite significant, and should be factored in when designing a cost-efficient wind set-up forecasting system for Lake Champlain.

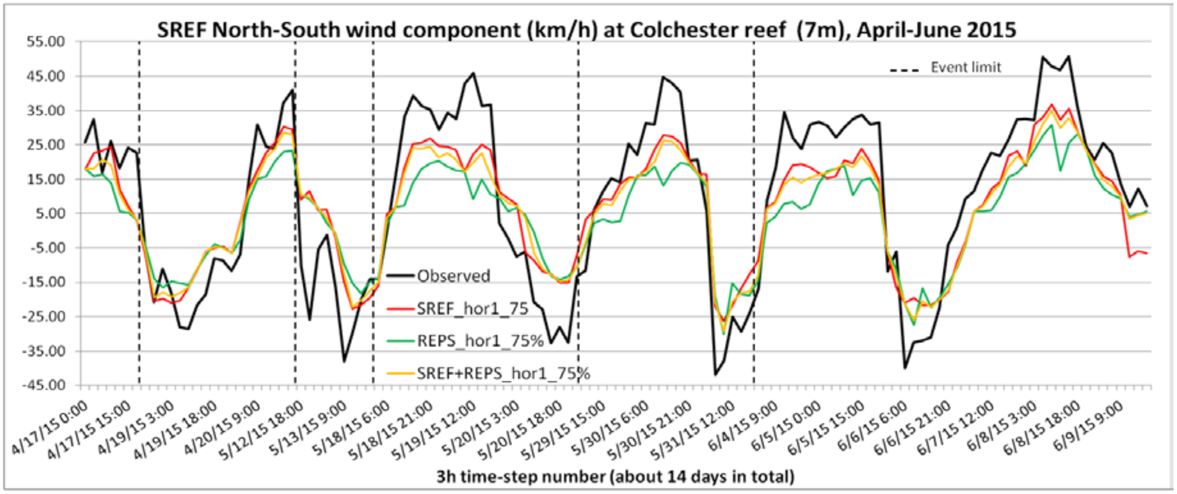


Figure 8: N/S wind component (km/h) at Colchester reef for the 2015 selected events (124 3-h time-step forecast/observation pairs). Forecasts on the graph are derived from the SREF, REPS, or merged product's (SREF+REPS, see text) third quartile (75%) and for the horizon 0-24h.

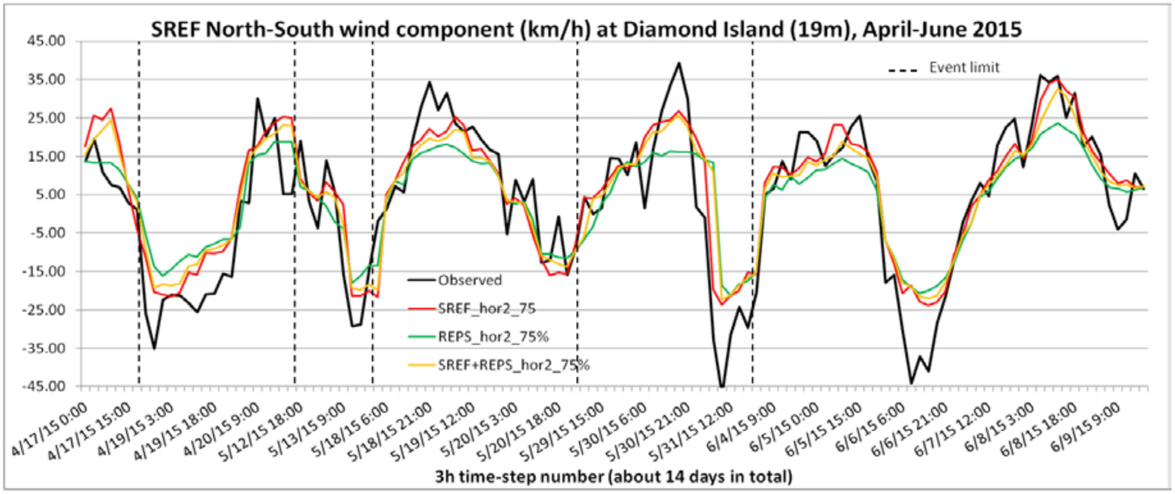


Figure 9: N/S wind component (km/h) at Diamond island for the 2015 selected events (124 3-h time-step forecast/observation pairs). Forecasts on the graph are derived from the SREF, REPS, or merged product's (SREF+REPS, see text) third quartile (75%) and for the horizon 24-48h.

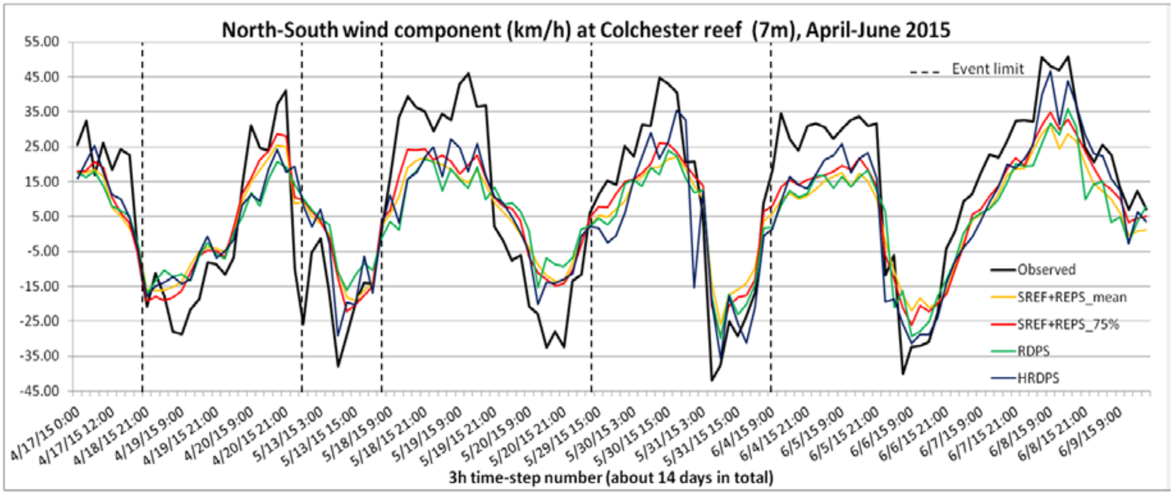


Figure 10: N/S wind component (km/h) at Colchester reef for the 2015 selected events (124 3-h time-step forecast/observation pairs). Forecasts on the graph are derived from the SREF + REPS merged product's (see text) mean or third quartile (75%), or from the RDPS or HRDPS models, and for the horizon 0-24h.

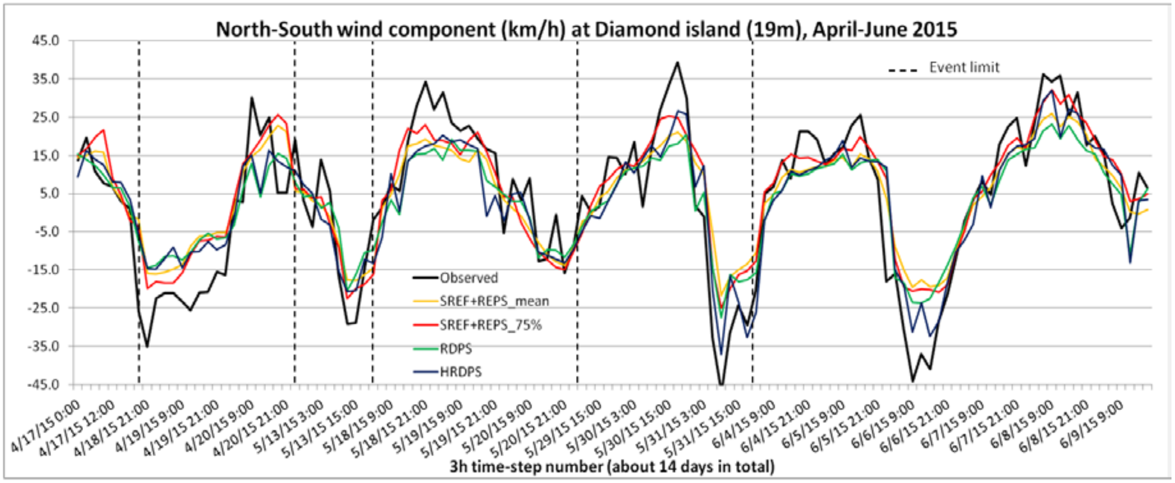


Figure 11: N/S wind component (km/h) at Diamond island for the 2015 selected events (124 3-h time-step forecast/observation pairs). Forecasts on the graph are derived from the SREF + REPS merged product's (see text) mean or third quartile (75%), or from the RDPS or HRDPS models, and for the horizon 0-24h.

Evaluation of precipitation forecasts

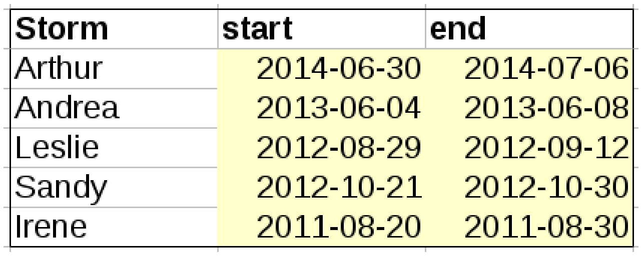
Although accurate surface wind forecasting require fairly high horizontal resolution in complex terrain, it is expected that long-lead precipitation forecasts based on models running at lower resolution can be useful for hydrological forecasting, especially for large scale storms such as extra-tropical cyclones. There are many reasons for this: (1) large-scale precipitation events have a long-term impact on water levels, whereas wind set-up events have little long-term impact on water levels, hence timing errors in a precipitation forecast are more forgiving than for a wind forecast; (2) location errors are also more forgiving for precipitation, since precipitation falling over the whole watershed of Lake Champlain have a lasting impact on water levels, whereas only wind events occurring over the surface of the lake impact water levels; and (3) whereas precipitation is a scalar, both wind speed and wind direction need to be correctly forecasted in order to be able to predict the impact of wind on water levels. On the other hand, precipitation is heavily parameterized in NWP models, which can cause significant biases in precipitation forecasts that are not necessarily reduced by simply increasing horizontal resolution.

Since timing and location errors are considered to be less critical for precipitation than for wind, we focus in this section on daily accumulations of precipitation (rather than 15-min averages for wind), and watershed averages (rather than station-specific locations).

Identification of events for precipitation forecast evaluation

In order to perform the evaluation, a set of rainy events corresponding to extra-tropical storms occurring in the Lake Champlain area were selected. Over the period of 2011-2014, based on storm tracks archived by the US National Hurricane Center, all five extra-tropical storms that affected the basin were identified (Table 13).

Table 13: Extra-tropical storm that affected the Lake Champlain watershed between 2011 and 2014, with start and end date of the corresponding precipitation events



Storm	start	end
Arthur	2014-06-30	2014-07-06
Andrea	2013-06-04	2013-06-08
Leslie	2012-08-29	2012-09-12
Sandy	2012-10-21	2012-10-30
Irene	2011-08-20	2011-08-30

Overall, a total of 42 days were considered, with a few of them showing very strong 24-h accumulation of precipitation.

Verification dataset for precipitation forecast evaluation

Two reference datasets were used to evaluate the precipitation forecasts, namely the U.S. Hydrological Rainfall Analysis Project (HRAP) and the Canadian Precipitation Analysis (CaPA).

HRAP data can be downloaded at http://water.weather.gov/precip/download.php. It consists of hourly accumulations of precipitation on a 4-km regular grid, which were summed up to obtain daily amounts. Real-time CaPA analyses can be downloaded at http://weather.gc.ca/analysis. Archived CaPA analyses are available from Environment Canada upon demand. Visit the web page http://collaboration.cmc.ec.gc.ca/cmc/cmoi/product_guide for more information on CaPA. It consists of six-hourly and twenty-four hourly accumulations of precipitation on a 10-km grid. The 24-h product was used in this study. The precipitation amounts from both sources were in good agreement, so the mean of both daily precipitation amounts was used as the observation to which the forecasts were compared.

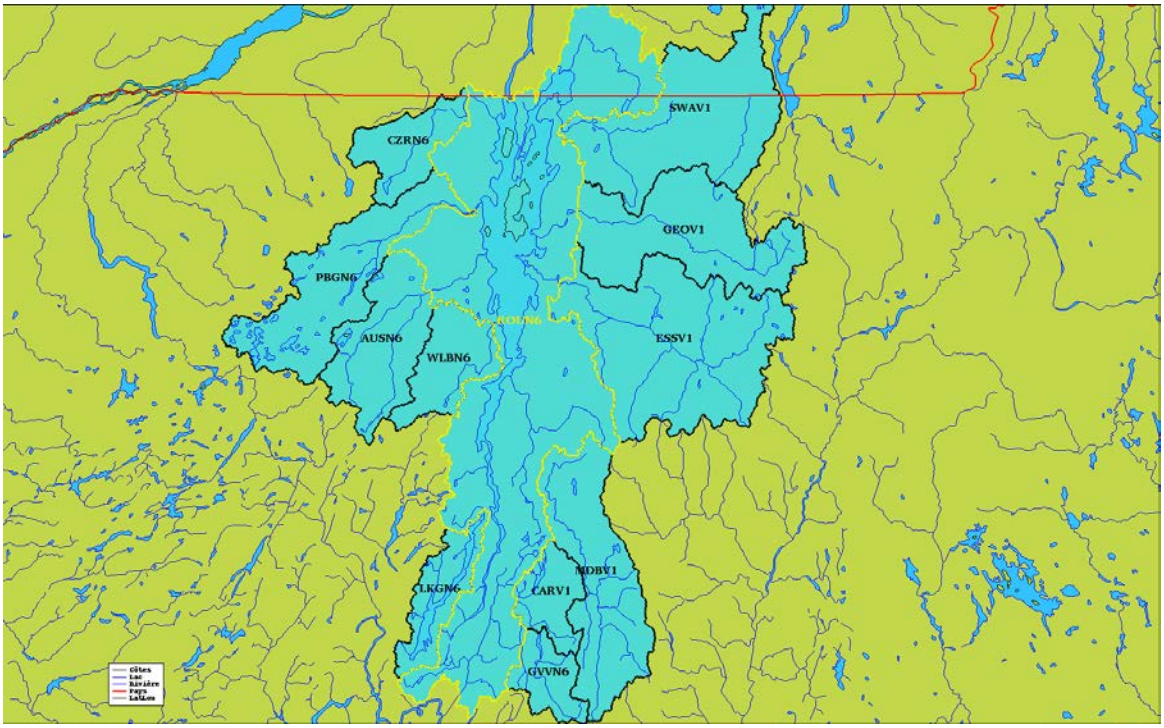


Figure 12: Sub-basins used for precipitation forecast evaluation

The entire watershed of Lake Champlain, subdivided into 12 sub-basins, was considered. Daily precipitation was computed for each of these sub-basins. These basins are identified on Figure 12. The Pearson correlation coefficient between HRAP and CaPA basin-averaged ranged from 0.93 and 0.99 depending on the sub-basin, confirming that the two products were very close to one another. It should however be mentioned that the correlation is largely driven by the largest event, which corresponds for all sub-basins to the maximum daily precipitation observed associated with Hurricane Irene. Figure 13 shows the scatterplot for the

42 days for the sub-basins which displayed respectively the lowest (LKGN6) and the highest (ESSV1) correlations. Not surprisingly, the former is one of the smallest watersheds, whereas the latter is one of the largest. It can be seen that the differences between HRAP and CaPA are important for LKGN6 for other events than the maximum observed precipitation; the correlation would be low if that event was excluded. In the case of sub-basin ESSV1, the linear fit is satisfactory for the smaller events as well.

Table 14: Correlation between daily precipitation estimated by HRAP and CaPA for the set of 42 days during which extra-tropical storms impacted precipitation over the Lake Champlain watershed between 2011 and 2014

Watershed ID	Correlation	Watershed ID	Correlation
AUSN6	0.98	**CARV1**	0.94
CZRN6	0.99	**ESSV1**	0.99
GEOV1	0.99	**GVVN6**	0.97
LKGN6	0.93	**MDBV1**	0.98
PBGN6	0.98	**ROUN6**	0.99
SWAV1	0.99	**WLBN6**	0.97

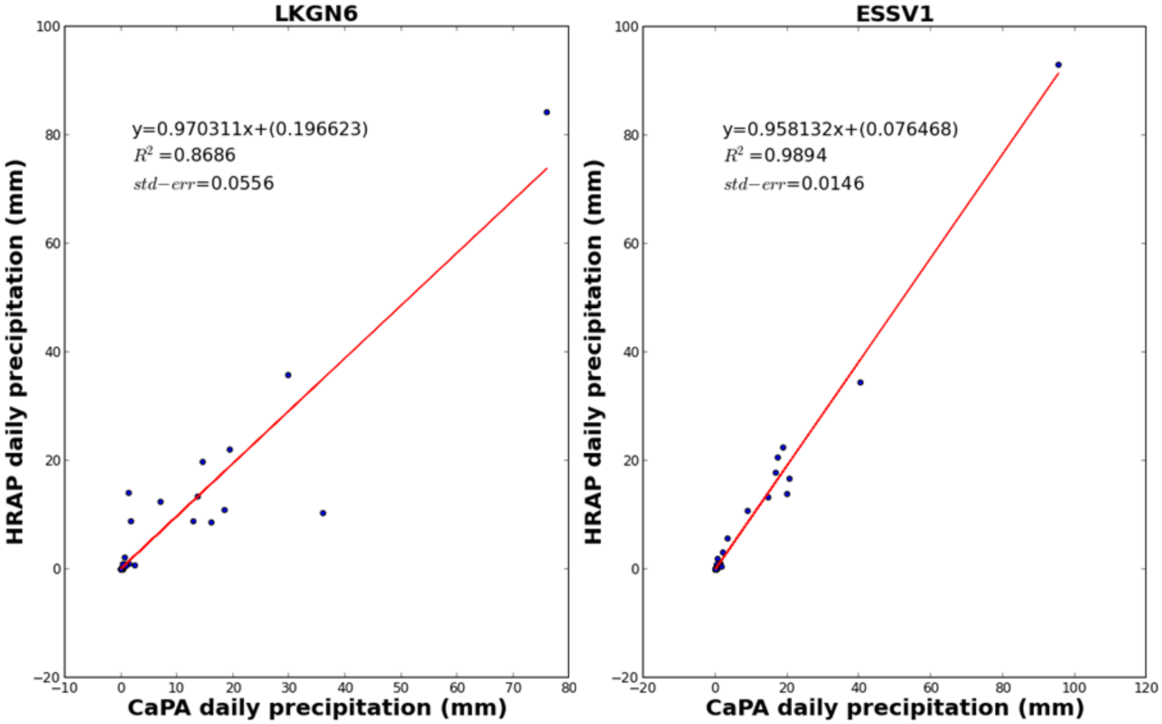


Figure 13: Scatterplot of CaPA vs HRAP daily precipitation for two sub-basins of Lake Champlain LKGN6 and ESSV1.

Precipitation forecast dataset

Canada, US and Mexico jointly participate in the North American Ensemble Forecasting System, which combines two-week forecasts from the Canadian Global Ensemble Prediction System (GEPS) and the NOAA Global Ensemble Forecasting System (GEFS). Both systems provide twenty scenarios (ensemble members) of daily forecasts, plus a control member, with forecasts being updated twice per day. Each ensemble member is obtained by perturbing initial conditions of the atmosphere and the surface in order to represent the observational and model uncertainties.

Focus was put on an evaluation of the GEPS forecasts, which have a horizontal resolution of 50-km, to assess how many days ahead it was possible to forecast precipitation with a deterministic forecast (the control member) and with an ensemble forecast (either using the ensemble mean or the whole distribution).

Precipitation verification results

In order to assess the skill of the precipitation forecast, the Pearson correlation coefficient was computed between the observed precipitation and the ensemble mean of the forecasted precipitation for each of the 12 sub-basins. The median of these 12 correlation coefficients was then obtained as a function of lead time. The same computation was also performed for the control member of the ensemble.

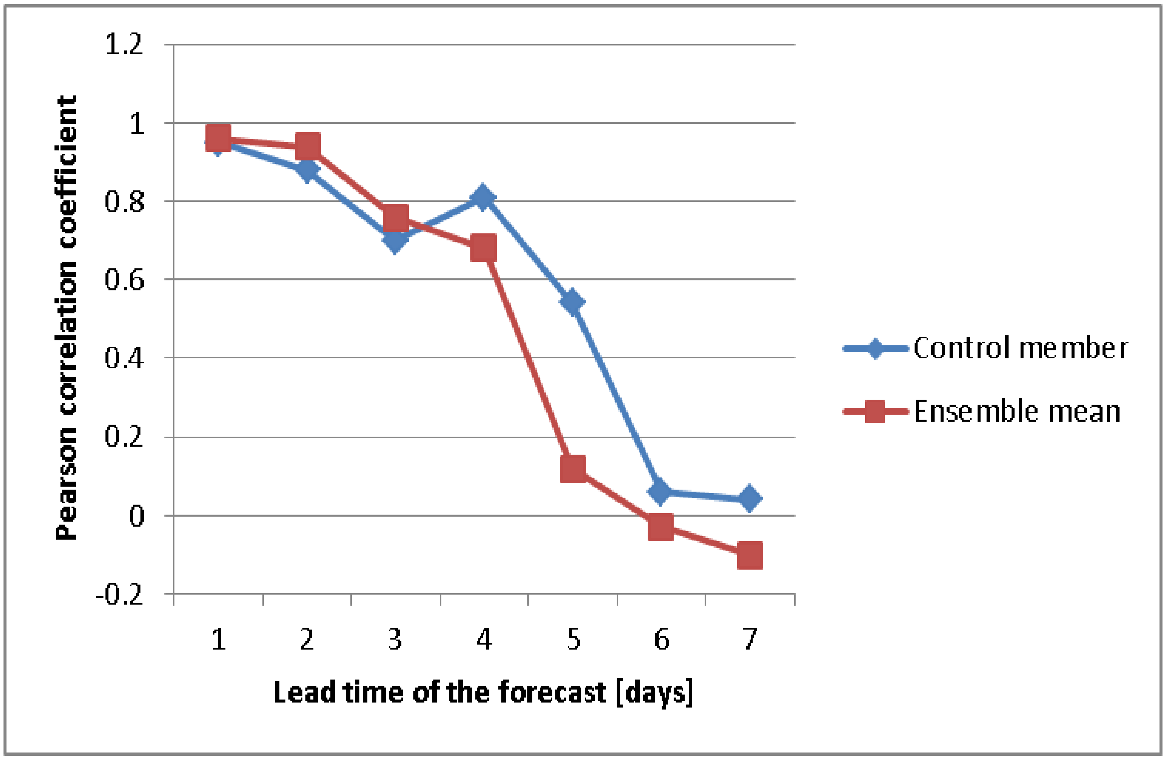


Figure 14: Median Pearson correlation coefficient between GEPS daily rainfall forecast and observed rainfall for twelve sub-basins of Lake Champlain

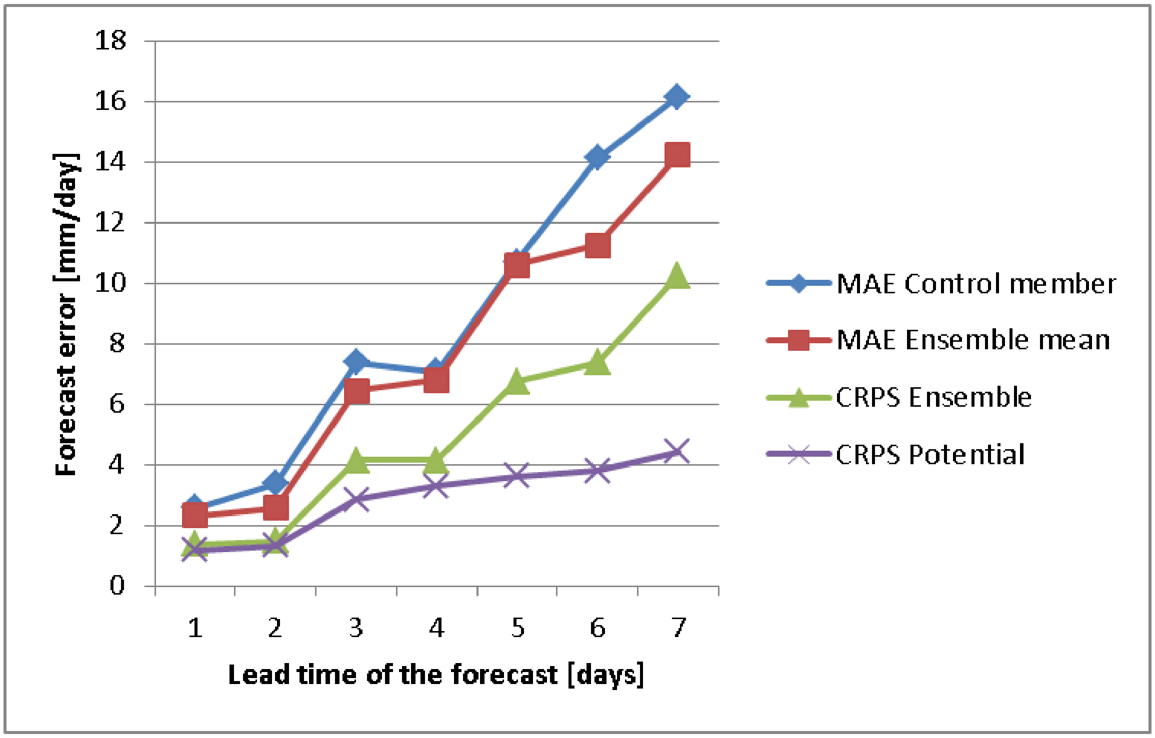


Figure 15: Median value of Mean Absolute Error (MAE) for GEPS control member and ensemble mean, as well as the GEPS Continuous Ranked Probability Score (CRPS) over twelve sub-basins used of Lake Champlain

Figure 14 shows how quickly correlation drops between forecasted and observed precipitation when considering either the control member or the ensemble mean of the GEPS. It can be seen that the ensemble mean correlates better with observed precipitation for the first three days of the forecast, but then drops quickly to zero over five days. The control member has slightly lower correlation for days 1-3, but then scores better than the ensemble mean for days 4-7. Correlation is however very small for days 6-7.

The MAE of the ensemble mean, the MAE of the control member and the CRPS of the ensemble were also computed for each sub-basin as a function of lead time, and in each case the median value over the 12 sub-basins was obtained and plotted (Figure 15).

For MAE, the ensemble mean outperforms the control member for all lead times, with differences being larger for days 6-7 when the correlation is zero with observed precipitation. The CRPS of the ensemble is always much smaller than the MAE of the ensemble mean, which suggests that there is useful information in the forecast uncertainty assessment provided by the ensemble members. Potential CRPS is also shown on Figure 15. The fact that it is almost equal to total CRPS for days 1-2 suggests that, contrary to wind forecasts, little can be done to easily obtain an improvement in performance through bias correction of the forecast. Although potential CRPS is much lower than total CRPS for longer lead times (days 5-7 especially), recall that the correlation of the ensemble mean with observed precipitation is small for these lead times.

Conclusion of precipitation forecast evaluation

In conclusion, forecasting precipitation for extra-tropical storm events seems possible up to five days in advance with some skill. The use of ensemble forecasts is however recommended in order to describe the uncertainty associated with that forecast. The evaluation of NAEFS forecast, which combines GEPS and GEFS members should also be performed in order to see if some skill could be recovered for longer lead times.

Hydrological and hydrodynamic models of Lake Champlain and the Richelieu River will likely require information on precipitation an time steps much finer than 24-h, likely on the order of 1-h. An evaluation of the hourly precipitation intensities forecasted by NAEFS ensemble members, which is outside the scope of this report, should therefore be performed before concluding that NAEFS forecasts can be used directly for hydrological or hydrodynamic forecasting.

Evaluation of degree-day forecasts

Temperature forecasts are generally more skillful than precipitation and wind forecasts, especially for long lead forecasts. In particular, temperature forecasts often have some skill even at a seasonal time scale, especially in winter. This study aims to see if the Canadian GEPS system shows skill for temperature more than a week into the future on Lake Champlain watershed. Because the objective is to contribute to a flood forecasting system, focus is put on weekly forecasts of basin-average degree-days of melt. Hence a forecast is considered perfect if it correctly estimates the average number of degree-days above zero observed over a week for the whole watershed. The variable of interest is thus:

$$X = \sum [\max(0,T_j^o) - \max(0,T_j^f)]$$

where T_j^o is the daily mean temperature for day *j*, T_j^f is the corresponding forecast, and the sum is computed over one week.

For lack of time, the same study could not be done with the GEFS. It is also not known whether the reforecast data required to perform the study exists for the GEFS.

Temperature forecast dataset

To assess skill of long lead forecasts, fairly large sample sizes are required. In order to meet this requirement, we rely here on a reforecast experiment performed with the GEPS: starting with ERA-Interim atmospheric reanalyses (Dee et al., 2011), GEPS forecasts are issued once a week from 1995 through 2012 (18 years) and integrated for four weeks. Horizontal resolution is 50-km. To limit computing time, the ensemble size of the GEPS reforecast was reduced from twenty to four members for the reforecast experiment. Given the small ensemble size, we only consider the ensemble mean as the forecast. Temperature forecasts are provided every six hours for those four weeks (at 00, 06, 12 and 18 UTC). Forecasts issued in February, March and April are considered, as these are the months during which snow melt typically occurs over the watershed.

Verification data and bias correction procedure

The forecasts are bias-corrected against ERA-Interim observed temperatures by removing the average bias observed for each synoptic hour over the 18 years: the same bias correction is used for all lead times and all forecasts, it only varies according to the hour of the day. The bias obtained for each synoptic time is presented in Table 15. All values are negative, which means that the forecasts are too cold, with 18 UTC forecasts (corresponding to the afternoon temperature) being almost unbiased. The negative bias is most important at night. Verification is done against this same temperature dataset. Although the same data is used both for bias correction and verification, the impact on scores is likely small given how simple the bias correction procedure is.

Table 15: Bias of temperature forecast for each synoptic hour

Synoptic hour	Local time	Bias of temperature forecast
00:00 UTC	20:00 ET	-1.77 C
06:00 UTC	02:00 ET	-2.57 C
12:00 UTC	08:00 ET	-1.22 C
18:00 UTC	14:00 ET	-0.01 C

Degree-day forecast evaluation

Figure 16 shows how the mean absolute error (MAE) of the forecast grows with time for forecasts issued in February, March and April. These are average values obtained from 1995-2012 (18 years) and over four forecasts per month (1 forecast issued each week), so a total of 72 forecasts in each case. For example, the top left plot shows the error in degree-day forecasts for each lead-time (from one week ahead to four weeks ahead), for all forecasts issued in February. As expected, forecast error grows with lead time. Notice that error is larger in March and April than in February. This is partly because there is more going on during these last two months: in February, temperature is generally below zero so that the total number of degree days of melt remains small. Errors are hence also smaller.

The bottom plots on Figure 16 compare the forecast error to a naive climatological forecast corresponding to the mean number of degree-days for that week. A skill score is obtained by computing the difference between the MAE of the naive forecast and the MAE of the model forecast, and then dividing by the MAE of the naive forecast. This skill score has a maximum value of one for a perfect forecast. A value of zero indicates a forecast which has no more skill than climatology. A negative value indicates a useless forecast.

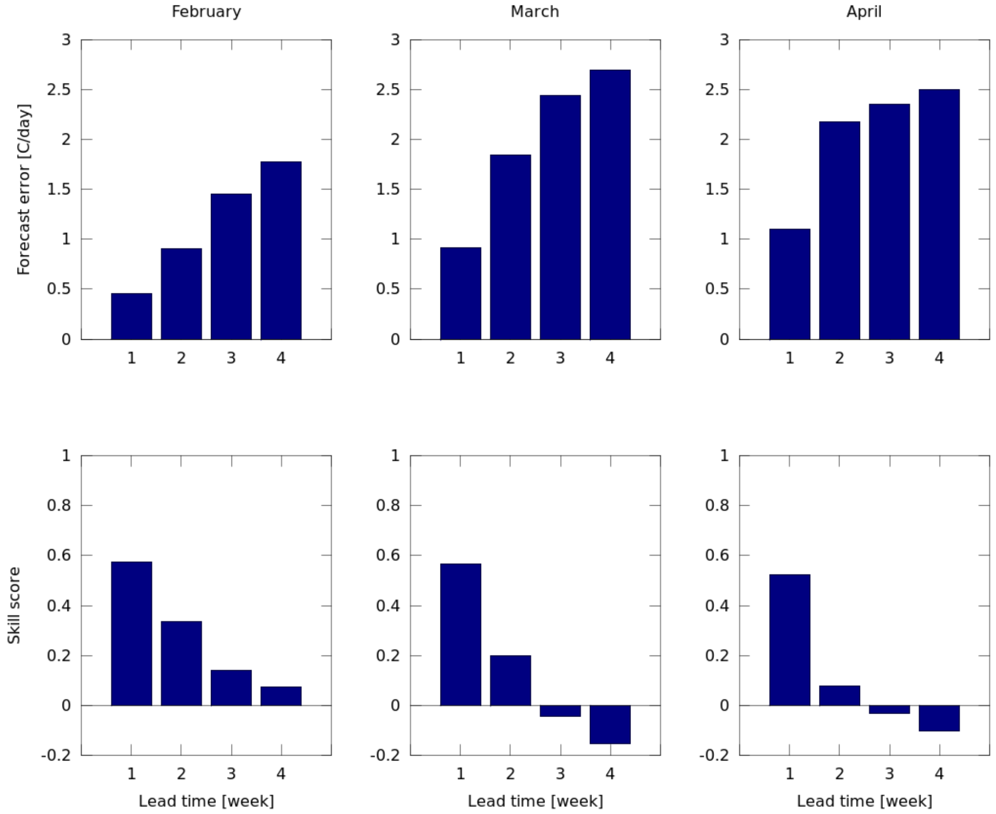


Figure 16: Mean absolute Error (MAE) and corresponding skill score of degree-day GEPS forecasts for Lake Champlain, from one to four weeks into the future.

According to these skill score plots, there is a similar level of skill for the first week for forecasts issued in all three months (above 0.5). For the second week, there is some skill for all three months but it is more modest and decreases from February to April. For weeks three and four, we only see skill for forecasts issued in February. March and April forecasts are useless beyond two weeks. Recall that February forecasts for weeks 3 and 4 are valid from the second half of February to the second half of March. That means that for later events such as the 2011 flood there is not a lot of useful skill after week 2.

This assessment is consistent with known characteristics of monthly and seasonal temperature forecasts for this region, where more skill is generally observed in winter. Whereas the use of monthly temperature forecasts might be warranted in February, for reliable flood forecasting later in the season it is probably more appropriate to rely on climatological temperature inputs for weeks 3 and 4. On the other hand, having the flexibility of using monthly ensemble weather forecasts for all three months could be useful in a warmer climate, since past observations of temperature become less representative of current conditions.

References

Côté, J., S. Gravel, A. Méthot, A. Patoine, M. Roch, A. Staniforth, 1998: The Operational CMC–MRB Global Environmental Multiscale (GEM) Model. Part I: Design Considerations and Formulation. Monthly Weather Review: Vol. 126, No. 6, pp. 1373–1395.

Dee, D. P., Uppala, S. M., Simmons, A. J., Berrisford, P., Poli, P., Kobayashi, S., Andrae, U., Balmaseda, M. A., Balsamo, G., Bauer, P., Bechtold, P., Beljaars, A. C. M., van de Berg, L., Bidlot, J., Bormann, N., Delsol, C., Dragani, R., Fuentes, M., Geer, A. J., Haimberger, L., Healy, S. B., Hersbach, H., Hólm, E. V., Isaksen, L., Kållberg, P., Köhler, M., Matricardi, M., McNally, A. P., Monge-Sanz, B. M., Morcrette, J.-J., Park, B.-K., Peubey, C., de Rosnay, P., Tavolato, C., Thépaut, J.-N. and Vitart, F. (2011), The ERA-Interim reanalysis: configuration and performance of the data assimilation system. Q.J.R. Meteorol. Soc., 137: 553–597. doi: 10.1002/qj.828

Girard, C., A. Plante, M. Desgagné, R. McTaggart-Cowan, J. Côté, M. Charron, S. Gravel, V. Lee, A. Patoine, A. Qaddouri, M. Roch, L. Spacek, M. Tanguay, P.A. Vaillancourt, and A. Zadra, 2014: Staggered Vertical Discretization of the Canadian Environmental Multiscale (GEM) Model Using a Coordinate of the Log-Hydrostatic-Pressure Type. *Mon. Wea. Rev.*, **142**, 1183–1196.

Hersbach, H. 2000. Decomposition of the Continuous Ranked Probability Score for Ensemble Prediction Systems. *Weather and forecasting*, 15: 559-570.

Matheson, J.E., Winkler, R.L. 1976. Scoring Rules for Continuous Probability Distributions. *Management Science* 22(10): 1087–1096.

Annex 1

Wind set-up events from 2011-2014 used for model forecast evaluation

event reference date	3/5/2011 0:00	3/7/2011 0:00	3/18/2011 0:00	4/17/2011 0:00	4/22/2011 0:00
start	3/4/2011 5:00	3/6/2011 9:00	3/18/2011 9:00	4/16/2011 2:00	4/21/2011 21:00
end	3/6/2011 6:00	3/8/2011 23:00	3/20/2011 2:00	4/18/2011 23:00	4/24/2011 6:00
duration (days)	2.0	2.6	1.7	2.9	2.4
mean wind (km/h)	37	20	21	17	16
mean dir (deg/N)	184			206	199
mean dir (deg/S)		178	171		
mean lev (m)	29.19	29.31	29.91	30.69	30.79
mean diff (cm)	4.9	-18.9	-13.5	7.6	9.4
max wind (km/h)	75.4	72.8	68.7	65.6	66.7
hour max wind	3/4/2011 13:00	3/7/2011 10:00	3/18/2011 12:00	4/17/2011 19:00	4/23/2011 13:00
max diff (cm)	8.1	-33.5	-19.0	25.8	33.0
Hour max diff	3/4/2011 18:00	3/7/2011 11:00	3/19/2011 5:00	4/16/2011 19:00	4/23/2011 15:00
event reference	5/10/2011 0:00	5/13/2011 0:00	5/17/2011 0:00	5/24/2011 0:00	6/11/2011 0:00
start	5/7/2011 16:00	5/12/2011 21:00	5/14/2011 19:00	5/21/2011 21:00	6/11/2011 1:00
end	5/12/2011 21:00	5/14/2011 17:00	5/18/2011 16:00	5/24/2011 17:00	6/13/2011 12:00
duration (days)	5.2	1.8	3.9	2.8	2.5
mean wind (km/h)	10	15	11	19	15
mean dir (deg/N)		186		195	185
mean dir (deg/S)	177		177		
mean lev (m)	31.33	31.17	31.20	31.25	30.75
mean diff (cm)	-6.5	6.4	-7.3	10.7	6.0
max wind (km/h)	42.8	42.2	47	58.8	75.3
hour max wind	5/10/2011 15:00	5/13/2011 12:00	5/15/2011 13:00	5/23/2011 12:00	6/11/2011 22:00
max diff (cm)	-25.5	13.6	-21.0	29.2	12.4
Hour max diff	5/9/2011 15:00	5/13/2011 14:00	5/15/2011 16:00	5/23/2011 15:00	6/12/2011 1:00
event reference	6/23/2011 0:00	7/18/2011 0:00	8/10/2011 0:00	8/23/2011 0:00	8/28/2011 0:00
start	6/23/2011 1:00	7/18/2011 0:00	8/9/2011 19:00	8/23/2011 0:00	8/27/2011 15:00
end	6/26/2011 20:00	7/19/2011 3:00	8/11/2011 13:00	8/26/2011 2:00	8/29/2011 3:00
duration (days)	3.8	1.1	1.8	3.1	1.5
mean wind (km/h)	14	11	15	19	17
mean dir (deg/N)	185	200	209	196	
mean dir (deg/S)					170
mean lev (m)	30.26	29.61	29.18	29.12	29.10
mean diff (cm)	6.4	6.4	8.1	14.6	-43.8
max wind (km/h)	45.8	34.5	64.2	67.3	84.1
hour max wind	6/25/2011 0:00	7/18/2011 13:00	8/10/2011 14:00	8/25/2011 2:00	8/28/2011 16:00
max diff (cm)	11.6	11.2	19.7	36.8	-116.4
Hour max diff	6/25/2011 1:00	7/18/2011 9:00	8/10/2011 1:00	8/24/2011 16:00	8/28/2011 16:00

event reference	10/17/2011 0:00	11/18/2011 0:00	12/15/2011 0:00	4/23/2012 0:00	6/11/2012 0:00
start	10/14/2011 13:00	11/18/2011 10:00	12/14/2011 7:00	4/21/2012 8:00	6/10/2012 23:00
end	10/18/2011 21:00	11/20/2011 15:00	12/16/2011 10:00	4/23/2012 20:00	6/13/2012 6:00
duration (days)	4.3	2.2	2.1	2.5	2.3
mean wind (km/h)	18	39	33	21	25
mean dir (deg/N)	214	193	188		176
mean dir (deg/S)				170	
mean lev (m)	29.79	29.33	29.26	29.27	29.42
mean diff (cm)	10.0	18.7	17.0	-22.6	14.1
max wind (km/h)	67.3	72	82.2	50.1	52.6
hour max wind	10/15/2011 22:00	11/20/2011 8:00	12/15/2011 10:00	4/23/2012 5:00	6/12/2012 14:00
max diff (cm)	28.4	38.0	48.2	-47.7	30.2
Hour max diff	10/16/2011 0:00	11/20/2011 7:00	12/15/2011 11:00	4/23/2012 4:00	6/12/2012 15:00

event reference	9/8/2012 0:00	11/7/2012 0:00	4/7/2013 0:00	4/19/2013 0:00	5/25/2013 0:00
start	9/5/2012 22:00	11/7/2012 4:00	4/6/2013 21:00	4/18/2013 6:00	5/24/2013 0:00
end	9/9/2012 12:00	11/9/2012 10:00	4/8/2013 2:00	4/20/2013 1:00	5/26/2013 20:00
duration (days)	3.6	2.3	1.2	1.8	2.8
mean wind (km/h)	16	18	30	35	19
mean dir (deg/N)	192		188	186	
mean dir (deg/S)		174			159
mean lev (m)	28.82	29.12	29.36	29.61	29.51
mean diff (cm)	12.5	-18.9	20.3	24.1	-25.8
max wind (km/h)	82.9	53.4	72.4	75.3	65.1
hour max wind	9/8/2012 16:00	11/8/2012 18:00	4/7/2013 12:00	4/19/2013 21:00	5/25/2013 16:00
max diff (cm)	50.6	-41.9	57.8	53.4	-54.1
Hour max diff	9/8/2012 13:00	11/8/2012 14:00	4/7/2013 14:00	4/18/2013 22:00	5/25/2013 17:00

6/25/2013 0:00	7/11/2013 0:00	9/21/2013 0:00	11/7/2013 0:00	4/4/2014 0:00	4/29/2014 0:00
6/23/2013 20:00	7/9/2013 6:00	9/20/2013 7:00	11/4/2013 18:00	4/2/2014 22:00	4/29/2014 0:00
6/25/2013 15:00	7/11/2013 2:00	9/22/2013 8:00	11/7/2013 10:00	4/4/2014 11:00	5/1/2014 17:00
1.8	1.8	2.04	2.7	1.5	2.7
10	15	26.6	30	12	18
190	188	168	159		164
				152	
29.99	30.36	29.63	28.90	29.18	30.50
7.1	5.7	18	17.5	-17.9	7.3
39.6	44.7	71.6	67	38.8	46.5
6/25/2013 14:00	7/10/2013 7:00	9/21/2013 11:00	11/6/2013 21:00	4/3/2014 16:00	4/30/2014 6:00
13.8	14.3	45.49	37.3	-22.5	15.7
6/25/2013 7:00	7/10/2013 12:00	9/21/2013 16:00	11/6/2013 22:00	4/3/2014 10:00	4/29/2014 19:00

event reference	5/9/2014 0:00	6/13/2014 0:00	7/3/2014 0:00	9/11/2014 0:00	10/22/2014 0:00
start	5/9/2014 0:00	6/11/2014 5:00	6/27/2014 21:00	9/8/2014 4:00	10/21/2014 6:00
end	5/10/2014 23:00	6/14/2014 7:00	7/3/2014 14:00	9/11/2014 19:00	10/24/2014 22:00

duration (days)	2.0	3.1	5.7	3.6	3.7
mean wind (km/h)	22	19	14	21	17
mean dir (deg/N)	182	161	183	168	
mean dir (deg/S)					165
mean lev (m)	30.45	29.81	29.67	29.00	28.73
mean diff (cm)	7.7	6.9	8.1	13.0	-18.7
max wind (km/h)	58	54.8	48.8	75.7	53.9
hour max wind	5/10/2014 14:00	6/11/2014 13:00	7/1/2014 18:00	9/11/2014 4:00	10/23/2014 2:00
max diff (cm)	16.5	12.6	20.4	54.6	-39.1
Hour max diff	5/10/2014 1:00	6/12/2014 10:00	7/1/2014 16:00	9/11/2014 8:00	10/23/2014 11:00

Wind set-up events from 2015 used for model forecast evaluation

events	1	2	3	4	5	6	7	8	9	10	11
start (UTC)	4/17/15 0:00	4/17/15 19:15	4/20/15 3:30	5/12/15 15:30	5/18/15 6:00	5/19/15 21:15	5/28/15 0:00	5/31/15 4:15	6/4/15 4:00	6/5/15 22:15	6/7/15 1:45
end (UTC)	4/17/15 19:00	4/20/15 3:15	4/21/15 0:00	5/15/15 2:00	5/19/15 21:00	5/21/15 4:30	5/31/15 4:00	6/1/15 0:00	6/5/15 22:00	6/7/15 1:30	6/9/15 19:00
duration (days)	0.8	2.3	0.9	2.4	1.6	1.3	3.2	0.8	1.7	1.1	2.7
mean wind (km/h)	24.2	20.5	30.4	26.4	33.9	30.7	27.2	28.0	29.0	23.5	26.9
main direction (N/S)	S	N	S	N	S	N	S	N	S	N	S
mean dir (deg/N)	192.1		186.0		195.8		200.2		193.2		191.2
mean dir (deg/S)		179.9		144.4		127.5		196.7		187.9	
max mean wind (km/h)	39.1	36.8	52.6	65.3	48.6	48.8	48.0	49.1	43.1	47.3	57.1
Hour max wind (UTC)	4/17/15 3:00	4/18/15 19:15	4/20/15 20:00	5/12/15 20:15	5/19/15 12:00	5/19/15 23:00	5/30/15 12:00	5/31/15 6:15	6/5/15 14:00	6/6/15 6:45	6/8/15 13:45