Final Report: Data Assimilation

Nicolas Weaver (5851572)
N.M.Weaver@student.tudelft.nl
David Plazas (5695767)
D.PlazasEscudero@student.tudelft.nl

May 1, 2024

Introduction

In this project, we study the application of two important techniques for data assimilation: the Kalman Filter (KF) and the Ensemble Kalman Filter (EnKF). In particular, we focus on studying the effectiveness to estimate and predict the sea level h and velocity u in the western Scheldt estuary (see Fig. 1 for reference), on normal conditions, as well as during the period during the cyclone Xaver of December 6 2013, also known as the *Sinterklaasstorm* in The Netherlands. The peak water level in Vlissingen was the second highest one over the last century.

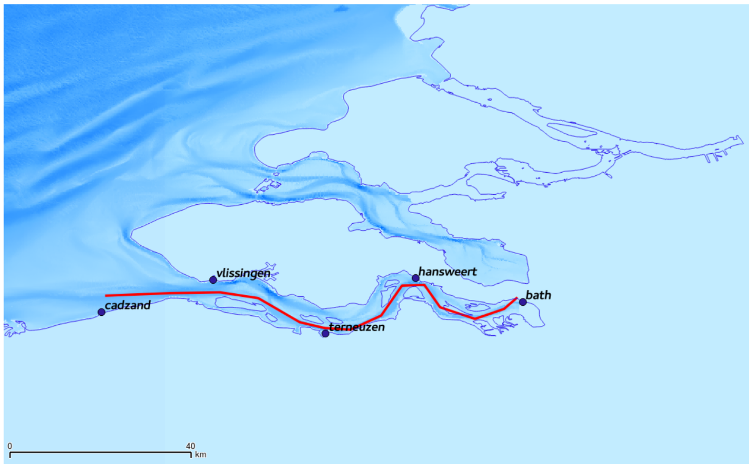


Figure 1: Western Scheldt estuary (image taken from project description).

Furthermore, we assume that we can model the quantities h and u by the linearized one-dimensional shallow water equations, also known as De St. Venant equations, which read

$$\begin{aligned} \frac{\partial h(x,t)}{\partial t} + D\frac{\partial u(x,t)}{\partial x} &= 0 \\ \frac{\partial u(x,t)}{\partial t} + g\frac{\partial h(x,t)}{\partial x} + fu(x,t) &= 0. \end{aligned} \tag{1}$$

Question 1

By definition, a function v follows a wave equation if and only if

$$\frac{\partial^2 v}{\partial t^2} = c^2 \frac{\partial^2 v}{\partial x^2}, \ c \in \mathbb{R}. \tag{2}$$

When the friction is neglected, the linearised shallow water equations become

$$\frac{\partial h}{\partial t} + D\frac{\partial u}{\partial x} = 0, \tag{3}$$

$$\frac{\partial u}{\partial t} + g\frac{\partial h}{\partial x} = 0, \tag{4}$$

and using Schwarz's theorem we have

$$\frac{\partial(3)}{\partial t} = \frac{\partial^2 h}{\partial t^2} + D\frac{\partial}{\partial x}\left(\frac{\partial u}{\partial t}\right) \tag{5}$$

$$= \frac{\partial^2 h}{\partial t^2} - Dg\frac{\partial^2 h}{\partial x^2}. \tag{6}$$

Likewise,

$$\frac{\partial(4)}{\partial t} = \frac{\partial^2 u}{\partial t^2} - Dg\frac{\partial^2 u}{\partial x^2}. \tag{7}$$

Therefore, by setting the propagation speed $c = \sqrt{Dg}$, the linearised equations become

$$\frac{\partial^2 h}{\partial t^2} = c^2\frac{\partial^2 h}{\partial x^2}, \tag{8}$$

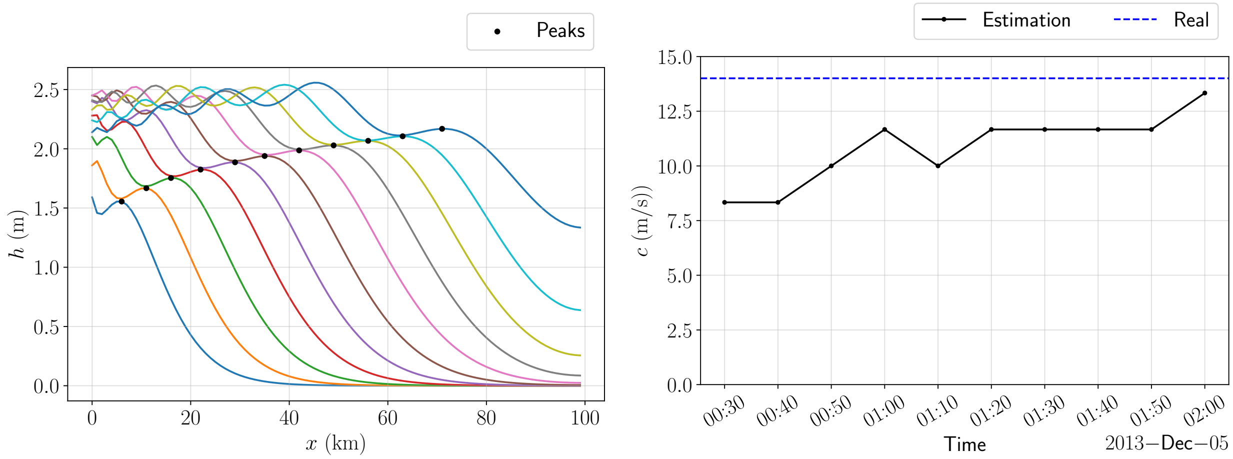
$$\frac{\partial^2 u}{\partial t^2} = c^2\frac{\partial^2 u}{\partial x^2}, \tag{9}$$

which are, by definition, wave equations. The numerical values of D in g are $D = 20.0\text{m}$ and $g = 9.81\text{m} \cdot \text{s}^{-2}$. So we can compute $c = 14.0\text{m} \cdot \text{s}^{-1}$.

Moreover, we can also try to estimate the wave propagation speed from the numerical simulations provided. Our approach follows an "intuitive" computation of speed: distance travelled divided by the time it took to travel said distance. For this, we attempted to follow a water blob through the domain, until it reached the right boundary. Around three-time steps, after the simulation started, we detected a local peak of the wavefront, which we selected as our reference point. This peak (local maxima) was then followed for the following time steps until it reached the right boundary. Since the time step is known, we are only left with how to estimate the distance travelled, but since the last local maximum was selected as the reference point, we can use a standard local maxima algorithm to find the last peak and check the position it is currently in.

The estimated peaks can be found in Fig. 2a, while the estimated velocities can be found in Fig. 2b. Note that, although the velocities vary from one estimation to the next one, they tend towards the theoretical estimation. However, in general, this approach depends heavily on the shape of the wave and visual aids as it is hard to estimate when a viable peak "enters" and "leaves" the domain. This can definitely be improved upon.

Another possible approach that was tried was using the wave equations (8)-(9) directly to estimate c. From the simulated data, we could try to estimate the second-order derivatives of h



(a) Estimated wavefront peaks (black dots) evolving through time.

(b) Estimated wavefront speed at each peak.

Figure 2: Wave front speed estimation results.

and u by a second-order finite difference centred at each point, and just solve for c at each point in the domain and time. This approach, however, led to numerical instability: the finite-difference approximation of the second derivative was highly inaccurate since the simulated data need not be smooth enough. In this sense, we preferred the more intuitive, yet biased, estimation.

Question 2

The De St Venant model that has been presented here is a simplified version of the shallow water equations, where linearity is assumed. However, a potential enhancement to this model would involve considering the non-linear aspects as well, which would provide a more accurate representation of the system under study.

From a numerical perspective, a higher level of precision can be achieved by employing a more advanced solver instead of the basic Euler method. For instance, utilizing a higher-order method like Runge-Kutta can yield more accurate results and ensure better convergence, or simply by taking smaller time steps.

Furthermore, improvements can be made in the measurements taken during the experiments or observations. By employing additional measurement tools, a more comprehensive dataset can be obtained, enabling a more detailed analysis of the phenomenon. Moreover, employing more precise materials for the experiments can also contribute to enhancing the accuracy of the measurements. However, it is worth noting that such improvements may require additional financial resources to acquire the necessary equipment or materials.

Question 3

The root-mean-square error (RMSE) and the bias are measures used to quantify the differences between model output (estimation) $\hat{\theta}$ and a "true" value θ, and they are given by

$$\text{RMSE}(\hat{\theta}, \theta) = \left(\mathbb{E}\left[(\hat{\theta} - \theta)^2\right]\right)^{\frac{1}{2}} = \left(\frac{1}{N}\sum_{i=1}^{N}\left(\hat{\theta}_i - \theta_i\right)^2\right)^{\frac{1}{2}} \tag{10}$$

$$\text{Bias}(\hat{\theta}, \theta) = \mathbb{E}\left[\hat{\theta} - \theta\right] = \frac{1}{N}\sum_{i=1}^{N}\left(\hat{\theta}_i - \theta_i\right) \tag{11}$$

We ran the model and compared the simulation results to the observations in the different stations. In Fig. 3, for each station, we have on the same plot, the observed data (in black) and the predicted data (in blue). We immediately notice that the simulation and the observation for the Cadzand station are almost the same. However, a few differences exist for the other stations.

In order to have a more quantitative comparison, we compute the RMSE and the bias for every station. The results are presented in Table 1. As expected the bias and the RMSE for the Cadzand station are the lowest. Moreover, the further the station is, the greater the bias is (in absolute value). The same can be said with the RMSE scores except for the Bath station. Nevertheless, the RMSE of those last four stations is similar in terms of order of magnitude.

Table 1: Bias and RMSE of h obtained at each station.

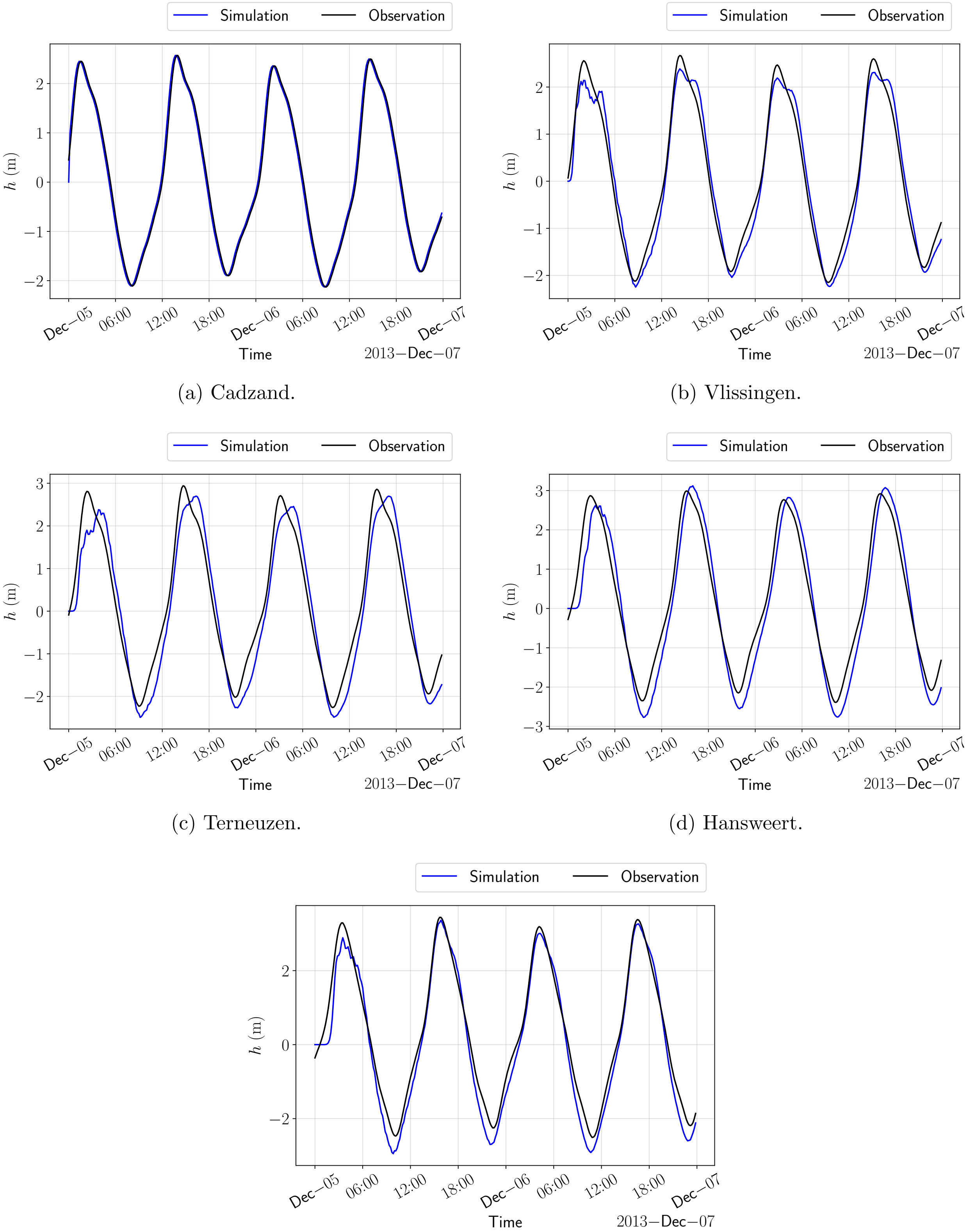
Station	**Bias**	**RMSE**
Cadzand	-0.0025	0.0418
Vlissingen	-0.029	0.3877
Terneuzen	-0.1302	0.6028
Hansweert	-0.1664	0.6045
Bath	-0.2452	0.4322

Other statistics exist to compute the efficiency of the model. For instance, we could have used the Mean Absolute Error (MAE). It is defined as:

$$\text{MAE}(\hat{\theta}, \theta) = \frac{1}{N}\sum_{i=1}^{N}\left|\hat{\theta} - \theta\right| \tag{12}$$

It is easy to interpret in a scatter plot. It is the average distance between each point and the ideal solution $\hat{\theta} = \theta$. The RMSE depends on the MAE but also on the variation of the errors and the square root of the number of points $\sqrt{N}$. For example, because the difference is squared, the RMSE gives a more important score to outliers than the RAE. Therefore the interpretation of the RMSE is more difficult [5].

The distinction between the MAE and the bias is that we sum absolute differences instead of algebraic differences. Therefore, a negative error can compensate for a positive error in the bias but not in the MAE. So the MAE gives a better view of the model's performance.



(a) Cadzand.

(b) Vlissingen.

(c) Terneuzen.

(d) Hansweert.

(e) Bath.

Figure 3: Water level measurements (black) and initial deterministic simulations (blue) at different locations.

Furthermore, the standard deviation is a commonly used statistic to measure the dispersion of the data. It is defined as

$$\sigma(\hat{\theta}, \theta) = \sqrt{\mathbb{E}\left[(\hat{\theta} - \bar{\theta})^2\right]} \tag{13}$$

The standard deviation is often used with the mean. If the standard deviation is low, then the values tend to be close to the bias. The RMSE score is a value that depends both on the bias and the standard deviation.

Another possible statistic is the Nash-Stutcliff Efficiency (NSE), originally proposed by Nash and Sutcliffe [4], which has been used in other hydrology applications. It is defined as

$$\mathrm{NSE}(\hat{\theta}, \theta) = 1 - \frac{N \cdot \mathrm{MSE}}{\sum_{i=1}^{N}(\hat{\theta} - \bar{\theta})^2}. \tag{14}$$

Note that its range of value is in $(-\infty, 1]$ with 1 being the optimal solution. If our predictions correspond to the mean of the true data, then $\mathrm{NSE} = 0$. This means that if the NSE of a prediction is negative, then taking the mean of the values is a better model. Unlike the bias and the RMSE, this criterion is normalised. Moreover, the comparison to the "mean" model is trivial with the NSE. Thus the interpretation is easier.

Question 4

Let $k \in \mathbb{N}_+$, and note that $\sigma_N^2(k) = \mathbb{V}\left[N(k)\right] = \mathbb{E}\left[N^2(k)\right]$. Assuming that the processes $N(k)$ and $W(k)$ are not correlated, the evolution equation yields

$$\begin{aligned}
\sigma_N^2(k+1) =& \mathbb{E}\left[N^2(k+1)\right] \\
=& \mathbb{E}\left[(\alpha N(k) + W(k))^2\right] \\
=& \mathbb{E}\left[\alpha^2 N^2(k) + 2\alpha N(k)W(k) + W^2(k)\right] \\
=& \alpha^2 \mathbb{E}\left[N^2(k)\right] + 2\alpha \mathbb{E}\left[N(k)W(k)\right] + \mathbb{E}\left[W^2(k)\right] \\
=& \alpha^2 \sigma_N^2(k) + \sigma_W^2.
\end{aligned}$$

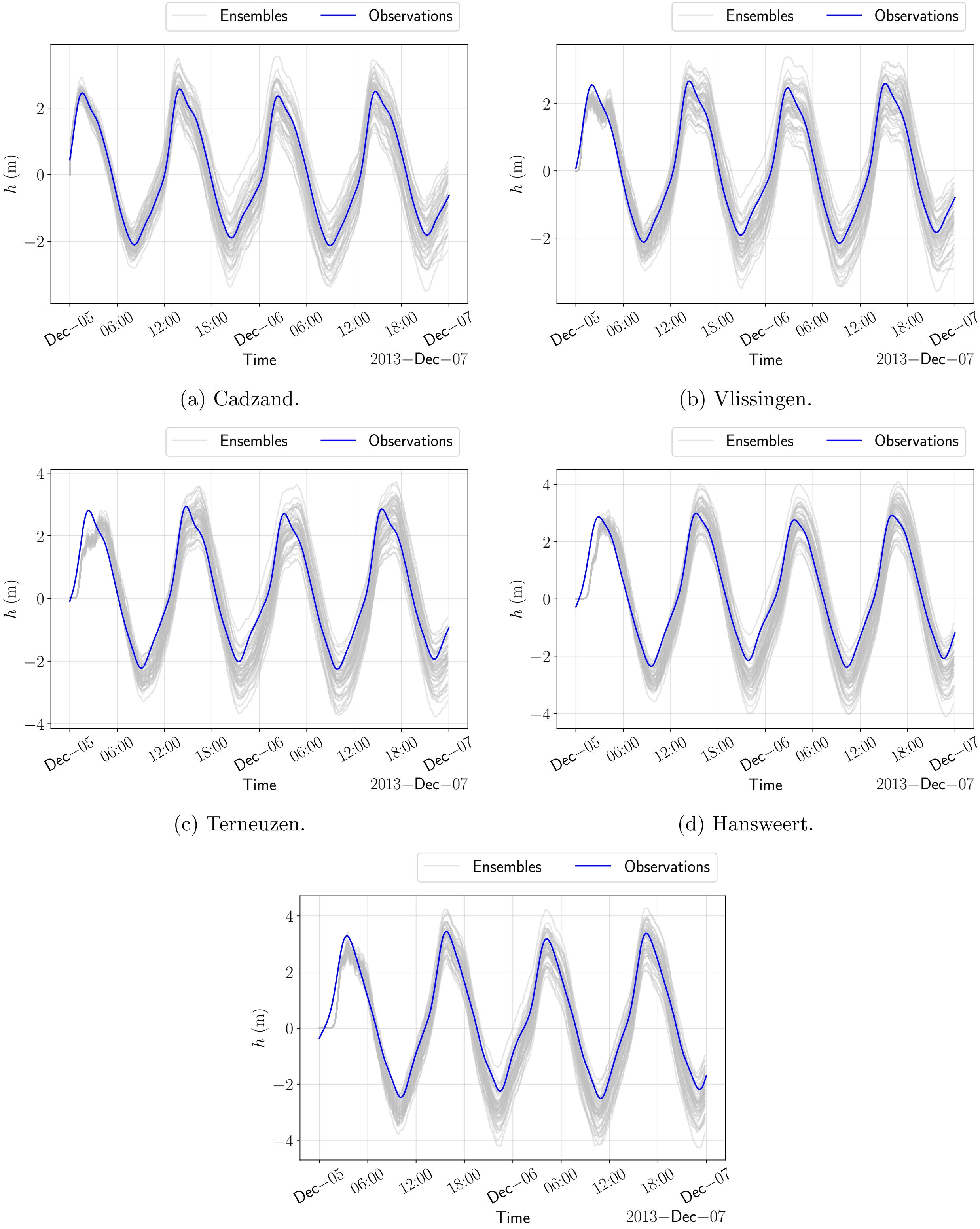
Therefore, for $k \to \infty$, assuming stationarity of the AR model, we get $\sigma_N^2 = \alpha^2 \sigma_N^2 + \sigma_W^2$, hence

$$\sigma_W = \sqrt{1 - \alpha^2}\sigma_N$$

We have $\alpha = e^{-\frac{dt}{T}}$ with $dt = 600s$ the timestep and $T = 6h$. So $\alpha = e^{-\frac{600}{6 \times 3600}} = e^{-\frac{1}{36}}$. Thus, in order to get $\sigma_N = 0.2$, we need to set

$$\begin{aligned}
\sigma_W =& \sqrt{1 - e^{-\frac{2}{36}}} \times 0.2 \\
\approx& 0.0465
\end{aligned}$$

The model was then run for $N = 50$ ensembles. The simulations, along with the observations of the height of the wave h at each station are presented in Fig. 4.



(a) Cadzand.

(b) Vlissingen.

(c) Terneuzen.

(d) Hansweert.

(e) Bath.

Figure 4: Water level observation of the simulated ensembles with stochastic boundary forcing at different locations (grey) and real measurements (blue).

The ensemble mean is the average value of the ensemble for a specified time. The ensemble spread is the standard deviation of the ensemble value with respect to the ensemble mean. In other terms, the ensemble spread is the RMSE of the ensemble mean [1]. In figure 5, the ensemble value is presented for both the height h and the velocity u at the observation stations. Note how the spreads seem to increase with time for the variable h at all locations. This means that the further in time we go, the more unpredictable the model gets. This is due to the chaotic behaviour of the system. On the long timescale, the system no longer depends on the initial conditions, whereas the model still relies on them.

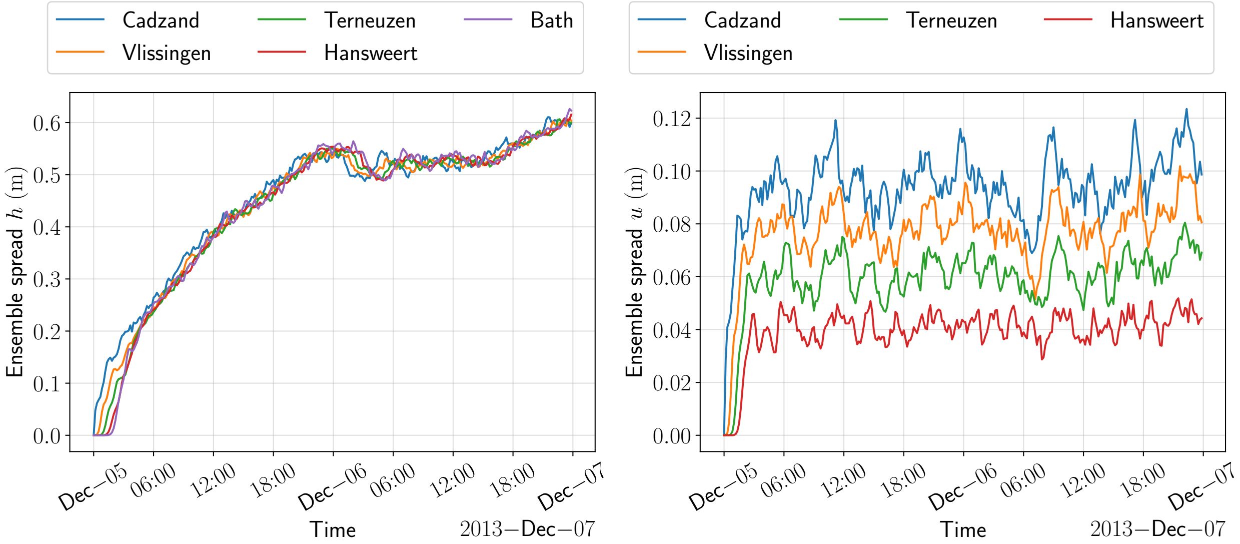


Figure 5: Ensemble spreads for both variables.

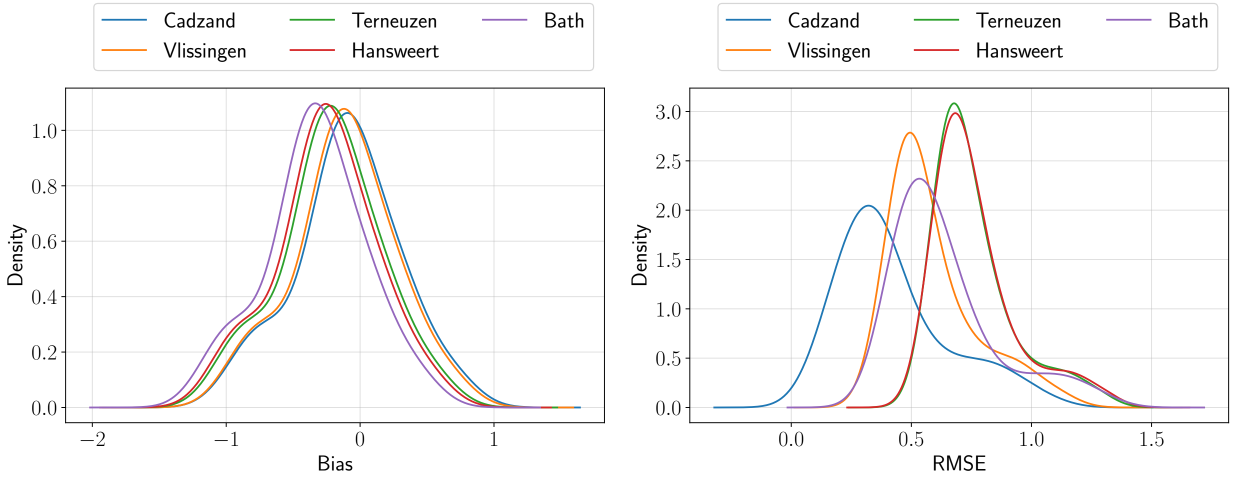
Finally, the RMSE and bias can be estimated through time, considering each ensemble as an estimator, and using the real observations as reference. In this way, one can obtain a series of RMSEs and biases, which are summarized using a histogram-like plot. In our case, we chose a Kernel density estimation for smoother and clearer results for comparison in a single plot. These results are presented in Fig. 6. We observe some similar results as in Fig. 3. First, the further the station the greater the bias is. Then, the RMSE score of the Cadzand station is significantly smaller than the others. Finally, the stations Vlissingen and Bath have a similar RMSE; so as the stations Teurneuzen and Hansweert.

Question 5

The Markov rule asserts that the new states should only depend on the previous state. First, the noise process is given by an AR(1) process

$$N(k+1) = \alpha N(k) + W(k),\ k \in \mathbb{Z}_+ \tag{15}$$

where $w \sim N(0, \sigma_W)$ and $\alpha \in \mathbb{R}$ as in the previous section. We can then clearly see that the future state $N(k+1)$ only depends on the present state $N(k)$ and noise $W(k)$. In the provided code (without taking into account the noise process), there is a function called `timestep(x, k,`



(a) Biases.

(b) RMSEs.

Figure 6: Density of the series of statistics at each measurement location.

`settings)`. The integer `k` is the time step index. `x` represents the vector state at the present time, namely, $\tilde{x}(k) \in \mathbb{R}^{200}$. In the `settings` script, we can find the state matrices $\tilde{A}$, $\tilde{B} \in \mathbb{R}^{200\times 200}$ and the boundary conditions $h_{\text{left}}(k) \in \mathbb{R}$. The boundary conditions are defined for every time k and in our state space representation will correspond to the input $\eta(k) \in \mathbb{R}$. The output of this function is the new state `newx` corresponding to $\tilde{x}(k+1)$.

Therefore, our evolution in time of the discretized system is described by the equations

$$\begin{aligned} \tilde{A}\tilde{x}(k+1) &= \tilde{B}\tilde{x}(k) + \tilde{C}\left(\eta(k) + N(k)\right), \\ N(k+1) &= \alpha N(k) + W(k),\ \ k \in \mathbb{Z}_+, \end{aligned} \tag{16}$$

but if we extend the state $x(k) := \text{col}\left(\tilde{x}(k), N(k)\right) \in \mathbb{R}^{201}$, we can obtain a more familiar state space representation with mass-matrix A as follows

$$Ax(k+1) = \tilde{M}x(k) + C\eta(k) + DW(k), \tag{17}$$

where A and $\tilde{M}$ are almost tri-diagonal matrices which have the form

$$A = \begin{pmatrix} \tilde{A} & 0 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 \\ * & * & * & 0 & \dots & 0 & 0 & 0 & 0 \\ 0 & * & * & * & \dots & 0 & 0 & 0 & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \dots & * & * & * & 0 \\ 0 & 0 & 0 & 0 & \dots & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 1 \end{pmatrix} \in \mathbb{R}^{201\times 201}$$

and

$$\tilde{M} = \begin{pmatrix} M & p \\ 0 & \alpha \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 1 \\ \ast & \ast & \ast & 0 & \dots & 0 & 0 & 0 & 0 \\ 0 & \ast & \ast & \ast & \dots & 0 & 0 & 0 & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \dots & \ast & \ast & \ast & 0 \\ 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & \alpha \end{pmatrix} \in \mathbb{R}^{201\times 201},$$

where the $\ast$ terms are the resulting coefficients from the discretization of the PDE, and they were omitted for simplicity, and $p = (1, 0, ..., 0)^T \in \mathbb{R}^{200}$ adds the forcing noise N to the boundary. Additionally, $C = (1, 0, ..., 0)^T \in \mathbb{R}^{201}$ fixes the boundary condition at the present state $x(k)$, and $D = (0, 0, ..., 1)^T \in \mathbb{R}^{201}$ adds the Gaussian noise to the AR state $N(k)$. It can be checked that this representation corresponds to system (17). Finally, we can take the inverse of matrix A to obtain the linear state space representation that we are familiar with

$$\begin{aligned} x(k+1) &= A^{-1}\tilde{M}x(k) + A^{-1}C\eta(k) + A^{-1}DW(k) \\ &= Mx(k) + B\eta(k) + GW(k) \end{aligned} \tag{18}$$

Note that the matrices M, B and G are time-independent, which will lead to a time-invariant Kalman filter. The new state $x(k+1)$ only depends on the present state $x(k)$, the present input $u(k)$ and the present noise $w(k)$. Thus, the model follows the Markov propriety.

Note also that the observations are modelled as

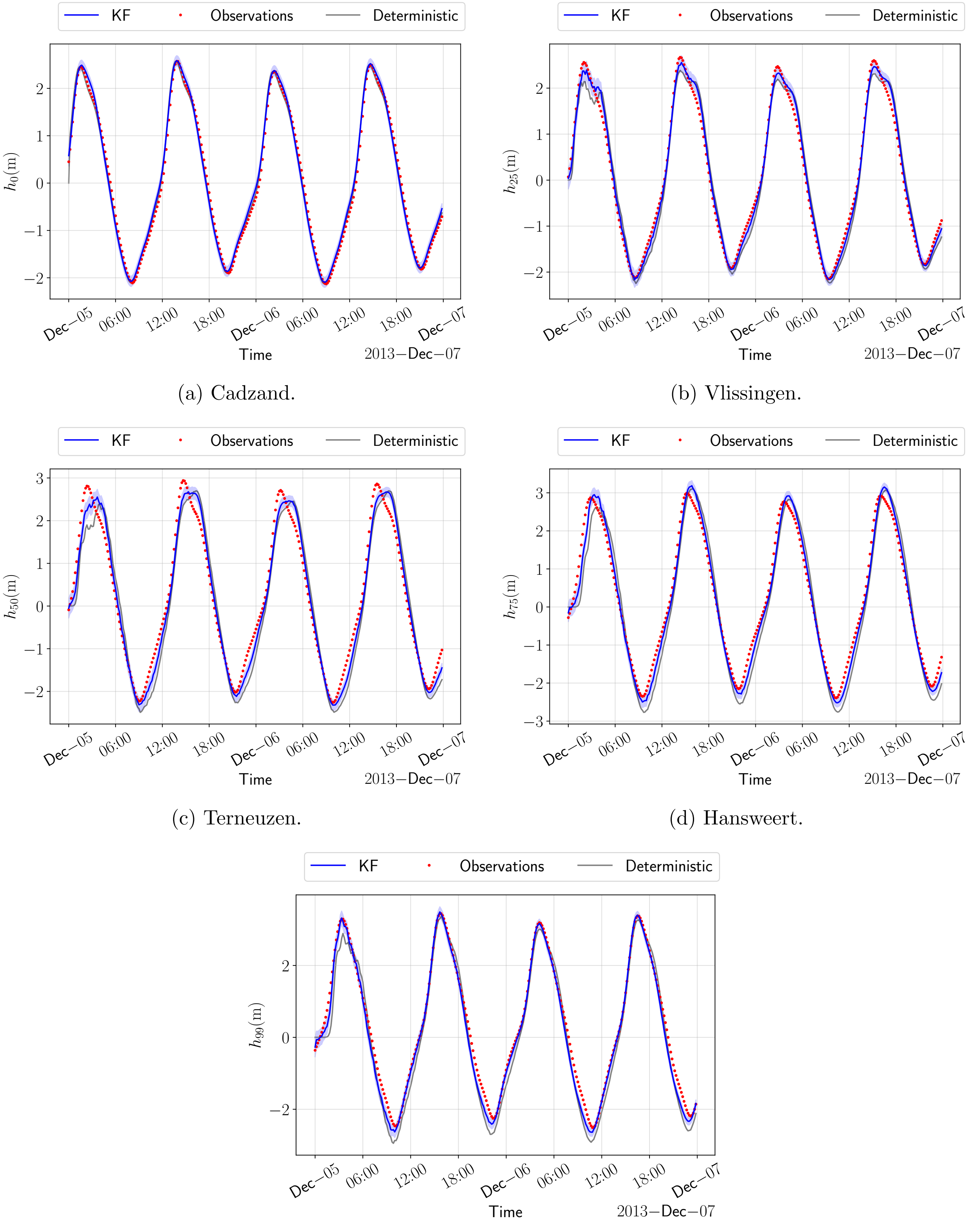
$$y(k) = Hx(k) + v(k),\ \ k \in \mathbb{Z}_+, \tag{19}$$

where $H \in \mathbb{R}^{5\times 201}$ extracts the water level at each measurement station (five in total), and $v(k) \sim N(0, R)$. Using representation (18), we can now apply the KF as usual (see [2, Sec. 27-28] for details on the KF) to obtain estimations of x based on observations y.

Thus, we ran a KF estimation setting the initial state as $x(0) = 0 \in \mathbb{R}^{201}$, the initial state covariance to $P(0) = 0.1 \cdot \mathbb{I}_{201}$ ($\mathbb{I}_n$ represents the identity matrix of dimension n), $R = 0.01 \cdot \mathbb{I}_5$. The estimated observations at the five measurement stations are presented in Fig. 7, and the obtained bias and RMSE at the five locations are presented in Table 2. Note that, overall, the KF yields good estimations of the state with relatively low bias and RMSE. Also, in Fig. 7, the deterministic simulations of the system were included as a reference for trying to model the system in a deterministic fashion, and we can observe that the KF estimations, up to a certain level, are between the deterministic simulations and the observations. This result agrees with what was expected since the KF can be interpreted as an interpolation between what is expected the model to do (deterministic simulation), and the observations themselves.

Table 2: Bias and RMSE of h obtained at each station.

Station	**Bias**	**RMSE**
Cadzand	0.0257	0.0814
Vlissingen	0.0168	0.2861
Terneuzen	-0.0575	0.4242
Hansweert	-0.0767	0.3928
Bath	-0.1431	0.2212



(a) Cadzand.

(b) Vlissingen.

(c) Terneuzen.

(d) Hansweert.

(e) Bath.

Figure 7: Estimation results using KF: estimated observations (blue), 1σ bands (light blue), real observations (red) and reference deterministic simulation (grey) at each location.

Question 6

Twin experiments are an important tool to test a data assimilation tool, as it provides access to the actual measurement and full state of the system. In our case, we decided to use a realisation of the system with stochastic forcing as our reference observation. Thus, the experiment consists of trying to replicate said realisation using an EnKF, which has an underlying model equal to the one that produced the reference data (hence the word "twin"). In these kinds of experiments, one can expect a perfect convergence of the EnKF to the actual data for two main reasons: first, as we will see later, the EnKF converges to the KF as $N \to \infty$, and it is well known that the KF guarantees that the resulting estimations are unbiased and of minimum variance (see [2, p. 470]), or a BLUE estimator as it is known in the literature; second, the as we used data from the same underlying system, you can argue that this realisation is "reachable" using the KF, as you would only need to adjust for the noise introduced in the model, but not the error induced by assuming a model simplified model of the real-world phenomenon.

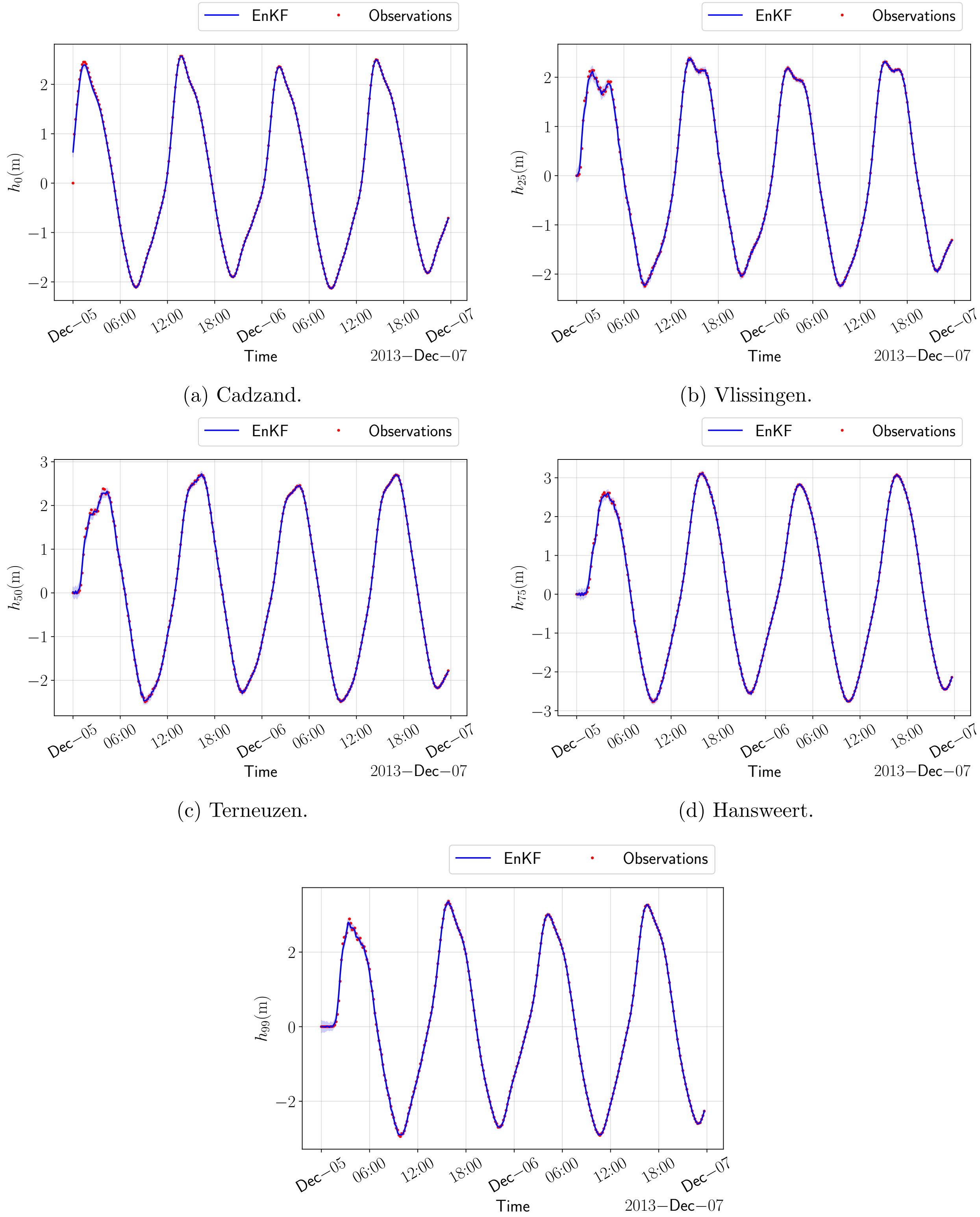
We performed a simulation using the ensemble Kalman filter with an ensemble size of $N = 50$. All of the other parameters are the same as in Question 5. The results of the identical twin experiment are presented in Fig. 8. We can see that the model fits perfectly with the observations. This is confirmed by the low values of the bias and the RMSE in Table 3.

Table 3: Bias and RMSE of h obtained at each station (twin experiment) for EnKF.

Station	**Bias**	**RMSE**
Cadzand	-0.0072	0.1381
Vlissingen	-0.0041	0.1406
Terneuzen	-0.0019	0.1496
Hansweert	-0.0006	0.1622
Bath	-0.0002	0.1743

Question 7

The EnKF should, in theory, converge to the KF (see [2, p. 541]), as our implementation makes use of the virtual observations. The expected rate of convergence is $\mathcal{O}\left(\frac{1}{\sqrt{N}}\right)$ [3]. We can execute the experiment for a growing number of N to verify the convergence rate. We can then compare the experimental convergence rate with the expected one. In Fig. 9 we compare the experimental error with the map $N \to \frac{5}{\sqrt{N}}$. We see that the curves fit. Thus we confirm the convergence rate.



(a) Cadzand.

(b) Vlissingen.

(c) Terneuzen.

(d) Hansweert.

(e) Bath.

Figure 8: Twin experiment estimation results using EnKF: estimated observations (blue), 1σ bands (light blue) and twin observations (red) at each location.

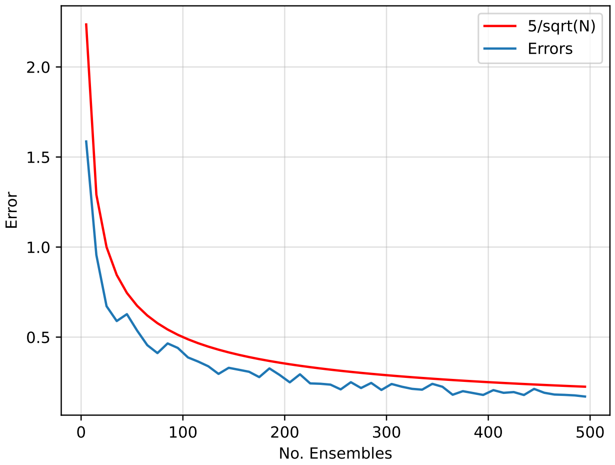
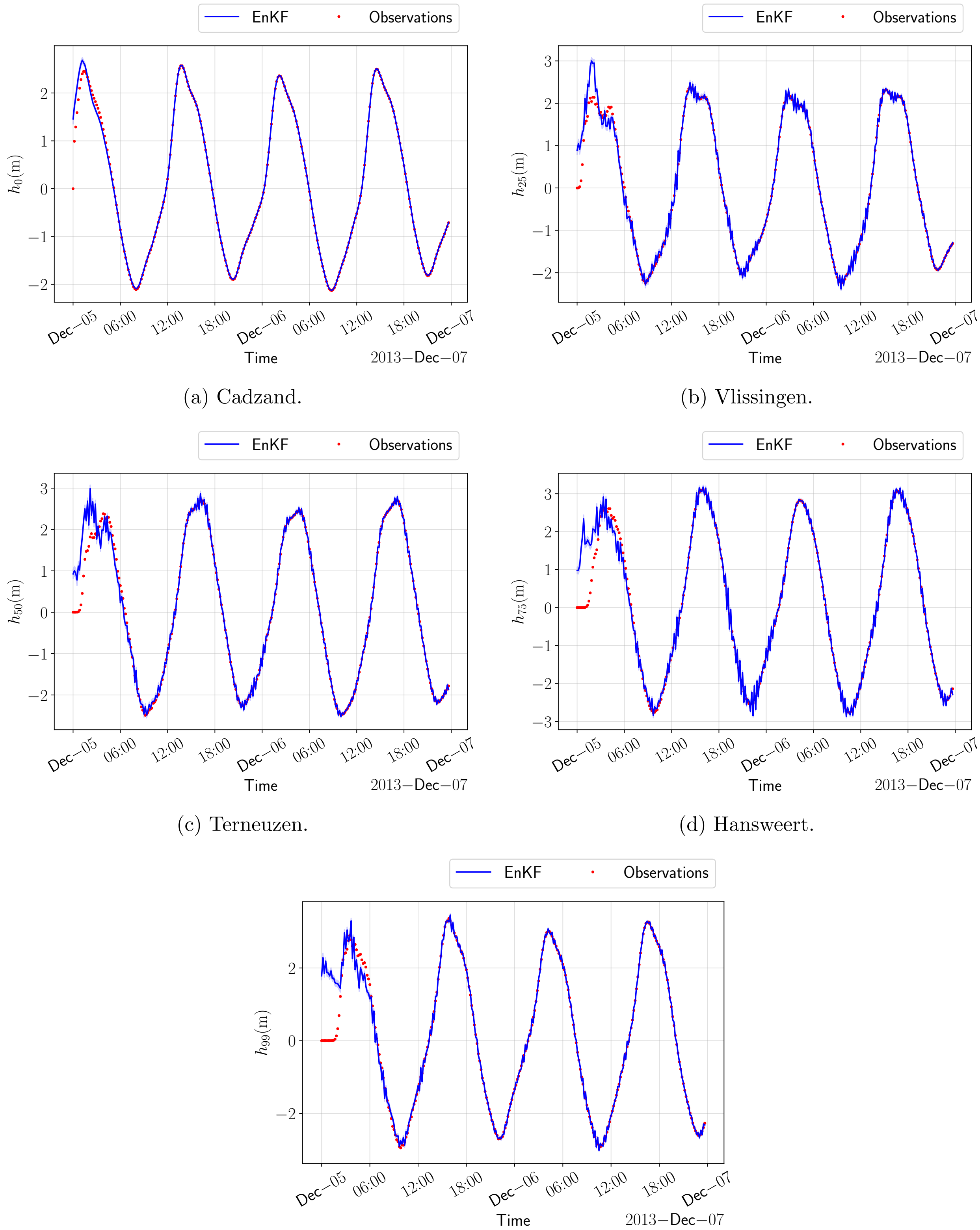


Figure 9: Convergence rate of the error (blue) and the representation of the map $N \to \frac{5}{\sqrt{N}}$ (red).

Question 8

For question 6 we chose $(h(x,0), u(x,0)) = (0,0)$ as initial conditions. We now choose $(h(x,0), u(x,0)) = (1,1)$ in their corresponding units. We then obtain the results presented in Fig. 10. We notice that on the short time scale, the EnKF and the observations are significantly different. This was expected as the chosen initial conditions differ from the observed initial conditions. However, after approximately six hours, the EnKF is able to match and track the observations. In other words, we have the same convergence despite the different initial conditions.

Furthermore, the difference in initial conditions results in outliers. Those have a big impact on the RMSE and are not relevant. Thus, the RMSE is no longer a pertinent statistic to analyse. However, the residuals $e(k) = \hat{y}(k) - y(k)$ are expected to decrease over time, as the KF corrects the state as more observations are available. The plot of residuals over time is presented in Fig. 11, note how all the residuals start away from zero, but eventually converge to a stationary distribution around zero, showing that the EnKF corrects the observations as time increases as well.



(a) Cadzand.

(b) Vlissingen.

(c) Terneuzen.

(d) Hansweert.

(e) Bath.

Figure 10: Twin experiment estimation results using EnKF: estimated observations (blue), 1σ bands (light blue) and twin observations (red) at each location, for the perturbed initial conditions.

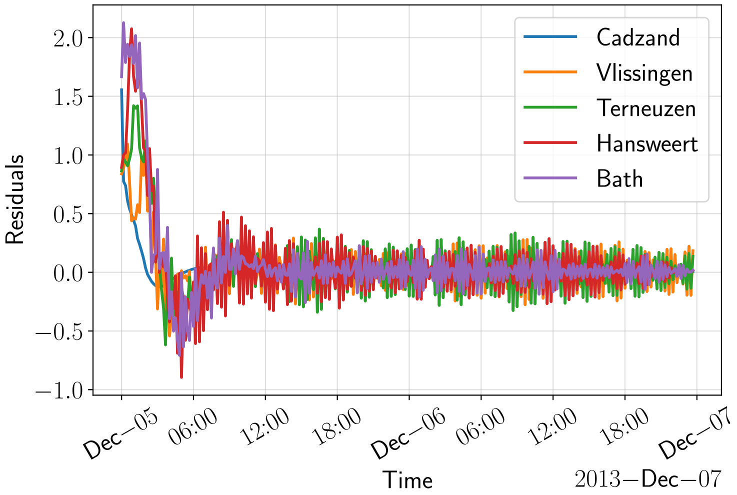
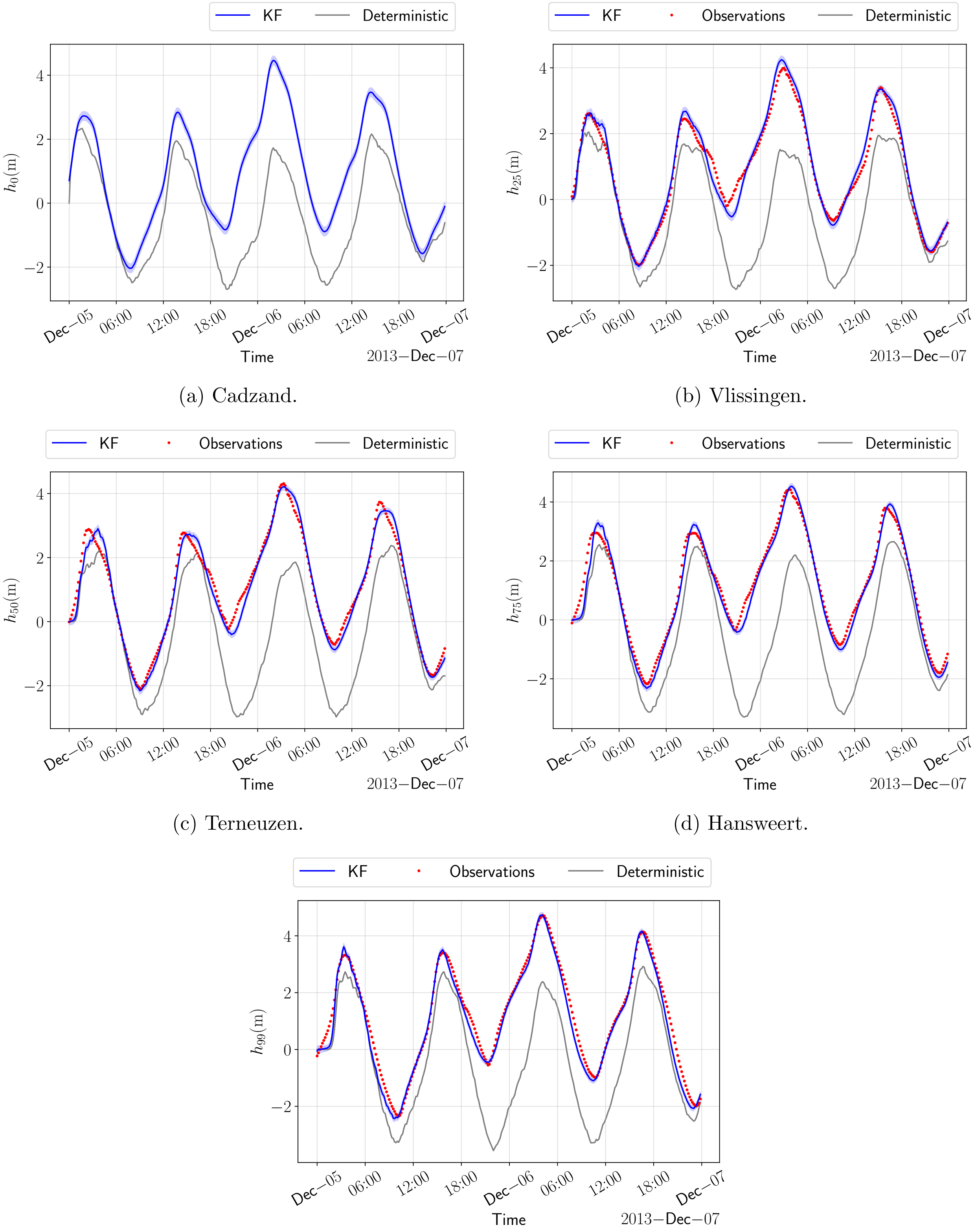


Figure 11: Residuals at each station of the EnKF model over time for perturbed initial conditions.

Question 9

In order to assimilate the cyclone observations, we modify our observation matrix H by eliminating our first row, since the boundary observations at Cadzand are missing.

Figure 12 presents the estimation of the storm with and without a Kalman filter at the five measurement stations. We also added the observations (except for Calzand) as red dots. We notice that, for all-time scales, the model using the Kalman filter is able to reproduce the observed data. On the contrary, the model without the Kalman filter is able to give a correct estimation only in the short time scale. On the long timescale, the differences are much more important. This difference was expected. Indeed, the Kalman filter has the ability to track the observation, unlike the deterministic model. So the error from the deterministic model increases with time.



(a) Cadzand.

(b) Vlissingen.

(c) Terneuzen.

(d) Hansweert.

(e) Bath.

Figure 12: Storm's estimation results using KF: estimated observations (blue), 1σ bands (light blue), storm measurements (red) and the model without a KF (grey) at each location.

Table 4 gives some statistics for the deterministic model. The time has been divided into four sets of 12h long. For each set, we compute the bias and RMSE for each station. As expected, the scores of the first set are better than the ones for the other sets. This confirms the fact that the deterministic model loses accuracy in time.

Table 5 gives the same statistics for the Kalman filter. We first notice that the Kalman filter has a lower RMSE and lower bias (in magnitude) than the deterministic model. Also, the values do not depend on time. Those statistics show similar results as in Fig. 12.

Table 4: Bias and RMSE of h obtained at each station for a four-time division for the deterministic model.

Stations		Vlissingen	Terneuzen	Hansweert	Bath
05/12 00:00 to 05/12 12:00	**Bias**	-0.285	-0.422	-0.533	-0.577
	RMSE	0.479	0.505	0.641	0.663
05/12 12:00 to 06/12 00:00	**Bias**	-0.620	-0.718	-0.742	-0.747
	RMSE	0.958	1.055	1.132	1.124
06/12 00:00 to 06/12 12:00	**Bias**	-1.475	-1.611	-1.675	-1.752
	RMSE	1.492	1.642	1.733	1.800
06/12 12:00 to 07/12 0:00	**Bias**	-0.714	-0.928	-1.075	-1.215
	RMSE	0.872	1.051	1.220	1.350

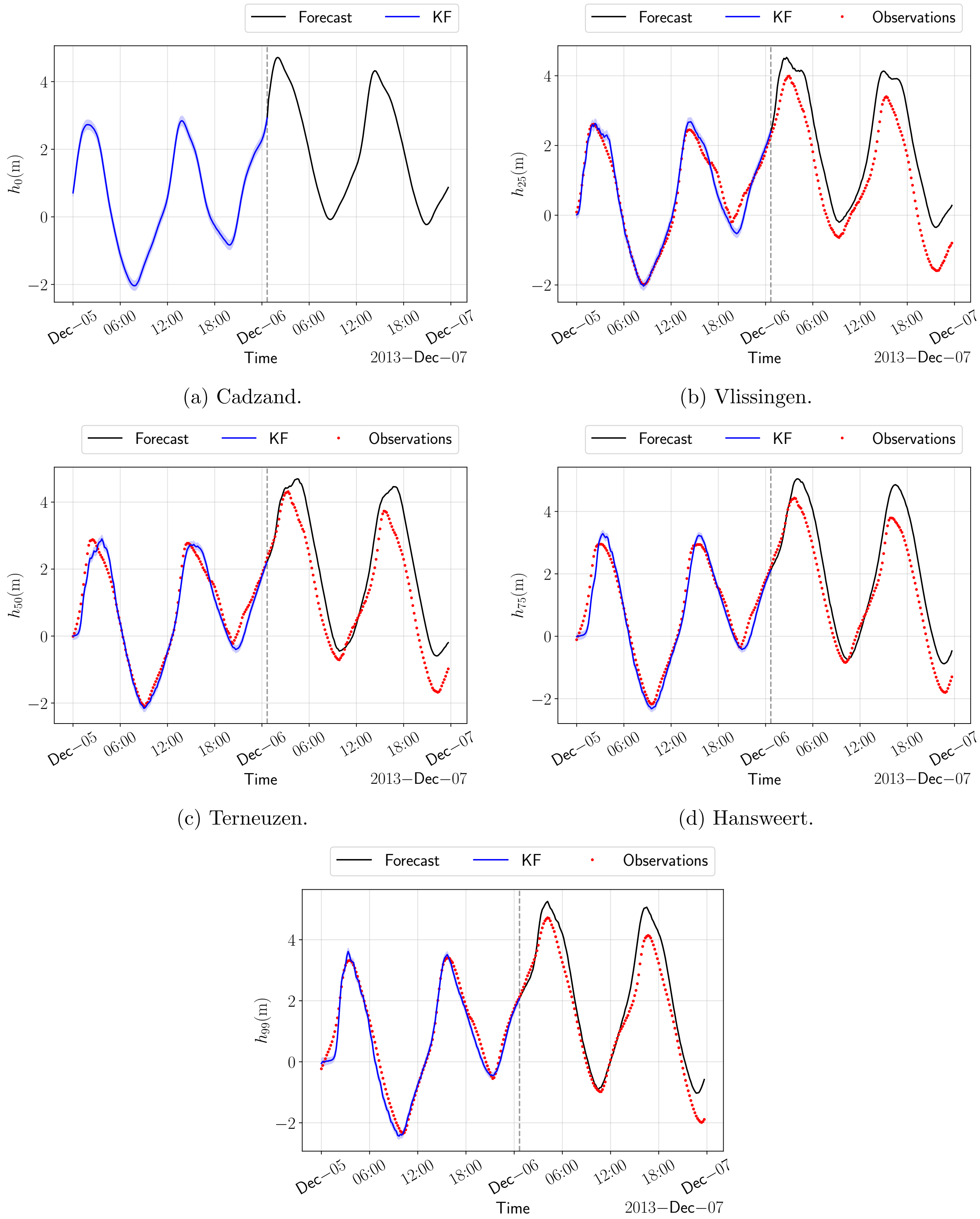
Table 5: Bias and RMSE of h obtained at each station for four time division for the Kalman filter.

Stations		Vlissingen	Terneuzen	Hansweert	Bath
05/12 00:00 to 05/12 12:00	**Bias**	0.047,	-0.082	-0.174	-0.203
	RMSE	0.517	0.249	0.262	0.300
05/12 12:00 to 06/12 00:00	**Bias**	-0.162	-0.205	-0.190	-0.171
	RMSE	0.701	0.382	0.239	0.225
06/12 00:00 to 06/12 12:00	**Bias**	-0.065	-0.075	-0.131	-0.193
	RMSE	0.514	0.214	0.160	0.235
06/12 12:00 to 07/12 0:00	**Bias**	0.166	-0.005	-0.102	-0.196
	RMSE	0.587	0.279	0.191	0.284

Question 10

The lead time is the time difference between the start of the forecast and the peak of the storm (i.e. maximum height). If the start of the forecast happens before the beginning of the storm, then the lead time is positive, otherwise, it is negative. Because the peak time is different according to the station, we will choose the peak time in Vlissingen as a reference. Our *a priori* expectation is that we obtain better results for shorter lead time.

We first fix the lead time at $l = 2h$ and have a look at all the stations. The situation is illustrated in figure 13. We observe that the the forecast tends to overestimate the measures. This can be confirmed by computing the biases. We can see in figure 15 that the biases for a lead time of two hours is indeed positive.



(a) Cadzand.

(b) Vlissingen.

(c) Terneuzen.

(d) Hansweert.

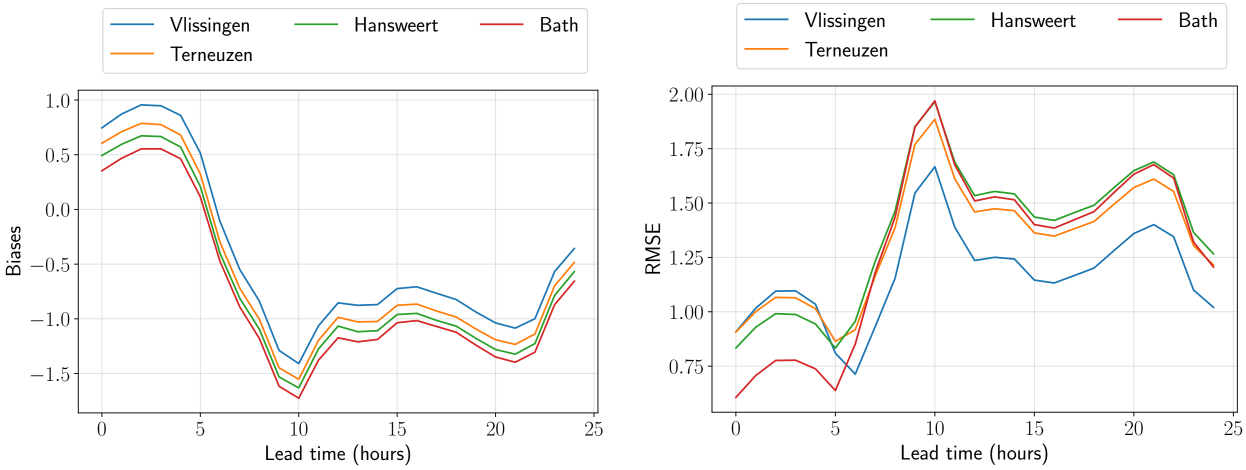
(e) Bath.

Figure 13: Forecasting results using KF: estimated observations (blue), 1σ bands (light blue) and storm measurements (red) at each location, for a lead-time $\ell = 2$h.

We then change the lead time and observe what appends. Figure 14 compares the forecast and the observation at the Vlissingen station for some different lead times. We notice that the lead has a lot of influence on the predictions. For some lead times, the forecast is overestimating and for some others, it is underestimating.

In order to better understand the impact of the lead time we compute the bias and the RMSE for every lead time from 0 to 24h (see figure 15). The statistics are calculated on the forecast dataset. This means that if the lead time is bigger, then the amount of data for the forecast is also greater. So the statistics of the forecast are not computed on the same database. We observe an optimal forecast for a lead time of 5h with a minimisation of the RMSE and a bias close to zero. For this lead time value, the start of the prediction is soon after the increase of the height.

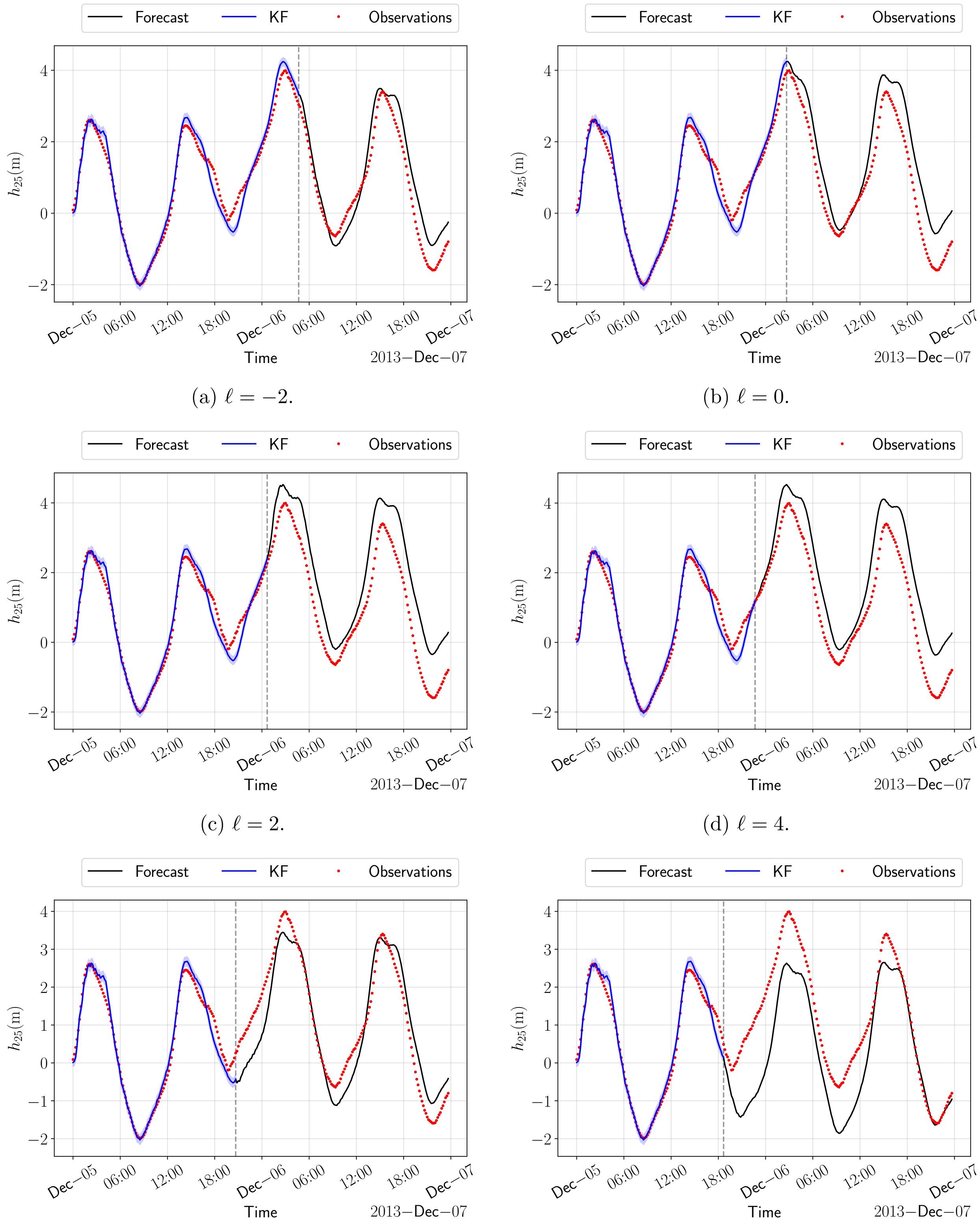
We also notice that the model performance is not decreasing as expected. But we can explain that. In Fig. 16 we have the forecast for lead time values of 10h and 15h. The main difference relies on the branch starting at 12:00 Dec 05 and ending at 17:00 Dec 05. This branch is part of the forecast for a lead time equal to 15h but not for the one equal to 10h. In this branch, the forecast is very accurate. So, on a global analysis, the statistics for the lead time of 15h are better than the ones for the lead time of 10h.



(a) biases

(b) RMSEs

Figure 15: Biases and RMSEs for different lead times



(a) $\ell = -2$.

(b) $\ell = 0$.

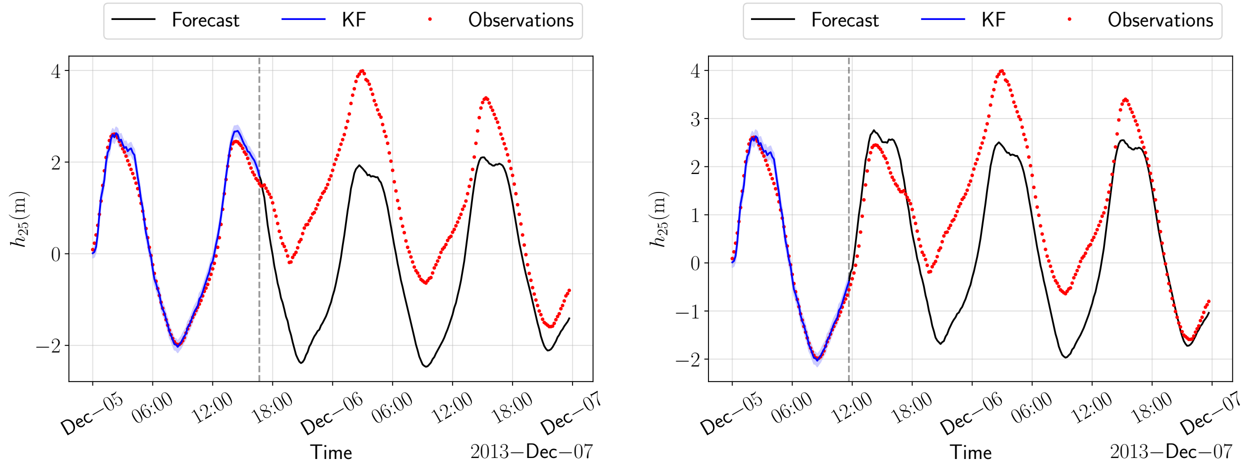
(c) $\ell = 2$.

(d) $\ell = 4$.

(e) $\ell = 6$.

(f) $\ell = 8$.

Figure 14: Forecasting results using KF: estimated observations (blue), 1σ bands (light blue) and storm measurements (red) at Vlissingen, for different lead-times ℓ.



(a) $\ell = 10$.

(b) $\ell = 15$.

Figure 16: Biases and RMSEs for different lead times

In real forecasting, we do not the lead time in advance as we do not know the real peak time. However, all the forecasts give the appropriate approximation of the peak time but not of the peak height. Therefore, for another experiment, we can use as peak times the ones corresponding to the beginning of the increasing slope. This will give different values of the peak height. We can then either choose to work with the interval of values or the mean value of this interval.

References

[1] URL: https://confluence.ecmwf.int/display/FUG/ENS+Mean+and+Spread.

[2] John M Lewis, Sivaramakrishnan Lakshmivarahan, and Sudarshan Dhall. *Dynamic data assimilation: a least squares approach*. Vol. 13. Cambridge University Press, 2006.

[3] Jia Li and Dongbin Xiu. “On numerical properties of the ensemble Kalman filter for data assimilation”. In: *Computer Methods in Applied Mechanics and Engineering* 197.43-44 (2008), pp. 3574–3583.

[4] J Eamonn Nash and Jonh V Sutcliffe. “River flow forecasting through conceptual models part I—A discussion of principles”. In: *Journal of hydrology* 10.3 (1970), pp. 282–290.

[5] Cort J Willmott and Kenji Matsuura. “Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance”. In: *Climate research* 30.1 (2005), pp. 79–82.