Convergence of Ensemble Simulations for Environmental Risk Assessment

Tor Nordam*, Ute Brönner, and Ragnhild L. Daae
SINTEF Materials and Chemistry
Trondheim, Norway
*tor.nordam@sintef.no

Abstract

Ensemble simulations of oil fate and transport for a hypothetical oil spill, sometimes referred to as "stochastic simulations", are frequently used for environmental risk assessment related to offshore operations. In this study we investigate the importance of the number of simulations and the effects of different sampling strategies on the results. We focus particularly on stranded oil, as this result is often assessed as a worst case criterion. We have run three ensembles for different discharge durations (3, 24 and 48 hours), all having the same simulation duration of 10 days. The ensembles were created by starting one simulation every hour throughout the four year period 2009 – 2012, resulting in a total number of 35064 simulations for each of the three durations. The complete set, as well as smaller subsets, were then used for further investigations.

We find large variations in the probability distributions for amount of oil in the environmental compartments atmosphere, ashore, sea surface, water column and sediments and for the biodegraded fraction. When calculating the autocorrelations of the series of simulation results, we found the shorter releases to have shorter correlation times. Based on these findings, we discuss the relationships between release duration, sampling rate and expected deviation caused by missing samples. Furthermore, we looked into sampling strategies, using both uniform and random sampling to create subsets from the complete set of simulations. We also briefly discuss the length of environmental data series to use for an environmental risk assessment, and briefly outline some future work.

1 Introduction

Ensemble simulation methods are commonly used in the forecasting of uncertain events, such as the weather. The Word Meteorological Organization's *Guidelines on Ensemble Prediction Systems and Forecasting* (WMO, 2012) has the following introduction:

> Ensemble Prediction Systems (EPS) are numerical weather prediction (NWP) systems that allow us to estimate the uncertainty in a weather forecast as well as the most likely outcome. Instead of running the NWP model once (a deterministic forecast), the model is run many times from very slightly different initial conditions. Often the model physics is also slightly perturbed, and some ensembles use more than one model within the ensemble (multi-model EPS) or the same model but with different combinations of physical parameterization schemes (multi-physics EPS).

The topic of the current paper is ensemble simulations, as used to investigate the possible outcomes of hypothetical oil spills, which is a common practice (see for example Price et al. (2003); Guillen et al. (2004)). When referring to an ensemble in this paper, we mean a collection of oil spill simulations, where the environmental forcing data varies between simulations, corresponding to the different initial conditions mentioned in the quote above. The variation in environmental data is achieved through varying the start date of the simulation, and selecting the corresponding data from an archive of hindcasts. For the ensembles described in the current study, all the hindcast data come from the same atmosphere and ocean models (see

Nordam, T., U. Brönner, and R.L. Daae, Convergence of Ensemble Simulations for Environmental Risk Assessment, Proceedings of the Thirty-ninth AMOP Technical Seminar, Environment and Climate Change Canada, Ottawa, ON, pp. 58-73, 2016.

Section 1.2 for details). An alternative approach would be to select hindcast data from among several archives generated by different atmosphere and hydrodynamic models, which would then be an example of what the WMO calls a multi-model Ensemble Prediction System (see for example Spaulding et al. (2014)).

The purpose of carrying out an ensemble of oil spill simulations can be to get information about possible outcomes, for example the likely sites of shoreline oiling, or it could be to statistically analyse the effect of response options (see for example Barker and Healy (2001)). Depending on the scenario of interest, the location of the spill might be known, while the time usually is not (although for an example of ensemble simulations applied to an ongoing spill, see Barker (2011)). The timing of an oil spill will determine the currents, winds, waves and other meteorological conditions, which together with the release parameters determine the transport and fate of the spill. Hence the common approach of carrying out large ensembles of oil spill simulations with different start times, using modelled historical data as environmental driving forces.

Since numerical simulations of oil spills are time consuming, one goal of an environmental risk analysis would typically be to get statistically robust results with the smallest possible number of simulations. While some work has been done to address the question of how to sample the environmental data (see for example Price et al. (2004)), clear guidelines seem to be lacking. The goal of the current study is to investigate the problem of how to utilise the available environmental data most efficiently, by which we mean to identify the smallest required number of simulations in order to achieve the required statistical robustness. Additionally, we discuss the required length of the environmental data series.

1.1 The OSCAR Oil Spill Model

For the oil spill simulations in this paper we have used OSCAR (Oil Spill Contingency And Response), which is a fully three-dimensional oil spill trajectory model for predicting the transport, fate and effects of released oil. The model accounts for weathering, the physical and chemical processes affecting oil at sea, as well as biodegradation. The development of models for these processes is strongly coupled with laboratory and field activities at SINTEF, on the fate and effects of oil and oil components in the marine environment (Brandvik et al., 2013; Johansen et al., 2015, 2013, 2003).

The OSCAR model computes surface spreading of oil, slick transport, entrainment into the water column, evaporation, emulsification and shore interactions to determine oil drift and fate at the surface. In the water column, horizontal and vertical transport by currents, dissolution, adsorption and settling are simulated. The different solubility, volatility, and aquatic toxicity of oil components are accounted for by representing oil in terms of 25 pseudo-components (Reed et al., 2000), which represent groups of chemicals with similar physical and chemical properties. By modelling the fate of individual pseudo-components, changes in oil composition due to evaporation, dissolution and biodegradation are accounted for. There is a biodegradation rate for each of the pseudo-components for the dissolved water fraction, droplet water fraction, surface and sediments.

OSCAR uses a pseudo-Lagrangian particle transport model, where the release is represented by numerical particles (Reed et al., 2000). Each model particle is tracked through the flow field, which is calculated from currents, wind, and ice if relevant. Buoyancy and sinking of oil droplets due to density differences or oil mineral aggregates are also included. The chemical composition of the released oil is an important part of the input to OSCAR. The crude oil considered in this study, Balder Blend (2010), has been characterised for use in OSCAR at the SINTEF oil lab.

1.2 The SINMOD Hydrodynamic Model

Current and wind data are required as input to the OSCAR model. The SINMOD hydrodynamic model was used to produce the current data (Slagstad and McClimans, 2005). SINMOD is based on the primitive Navier-Stokes equations and is established on a z-grid, using a constant-depth discretisation. The vertical turbulent mixing coefficient is calculated as a function of the Richardson number, Ri, and the wave state. The flow becomes turbulent when Ri is smaller than 0.65 (Price et al., 1986). Near the surface, vertical mixing due to wind waves is calculated from wind speed and fetch length. Horizontal mixing is calculated according to Smagorinsky (Smagorinsky, 1963).

The SINMOD model area used to generate the hydrodynamic data for this study is shown in dashed outline on the inset globe in Figure 1. The model area has a spatial resolution of $4\,\mathrm{km} \times 4\,\mathrm{km}$, and the dataset produced has a temporal resolution of 2 hours. Boundary conditions were taken from a larger model domain, at $20\,\mathrm{km} \times 20\,\mathrm{km}$ resolution. A total of 8 tidal components were imposed by specifying the various components at the open boundaries of the large-scale model. Tidal data were taken from TPXO 6.2 model of global ocean tides (Egbert et al., 2004)*. Wind and air temperatures have to be provided as input to SINMOD, and were in this case taken from the ERA-Interim Reanalysis (Dee et al., 2011).

2 Scenario and Simulation Setup

The base scenario used in this study is a surface release with a total released amount of 12000 metric tons of oil. This is a realistic amount for oil spills during for example loading from platform to tanker or a release following a ship collision. We have considered three versions of the same base scenario, with spill durations of 3, 24 and 48 hours. Since the total release amount is constant, the release rate is higher in the shorter releases. The main scenario parameters are described in Table 1, and the release location is shown on a map in Figure 1.

The simulations in this study were carried out as an ensemble of individual OSCAR simulations. This allows the results of each simulation to be considered independently, which in turn enables us to sub-sample the complete set of simulations in order to investigate convergence. Each simulation took approximately 450 seconds, which gives a total of about 13000 CPU hours for the three complete ensembles.

Table 1 Simulation parameters for ensemble simulations. Other than release duration and release rate, parameters were kept constant between the three ensembles.

Location	7.01 E, 63.50 N
Release amount	12000 metric tons
Release duration	3 h, 24 h and 48 h
Release depth	Surface release
Oil type	Balder Blend 2010
Simulation duration	10 days
Number of particles	5000 (dissolved) + 5000 (droplets)
Grid size	336× 278 cells
Cell size	1 km × 1 km
Simulation timestep	10 minutes
Environmental data	2009 – 2012, 4 km ×4 km, 2 hours

*See also `http://volkov.oce.orst.edu/tides/global.html`

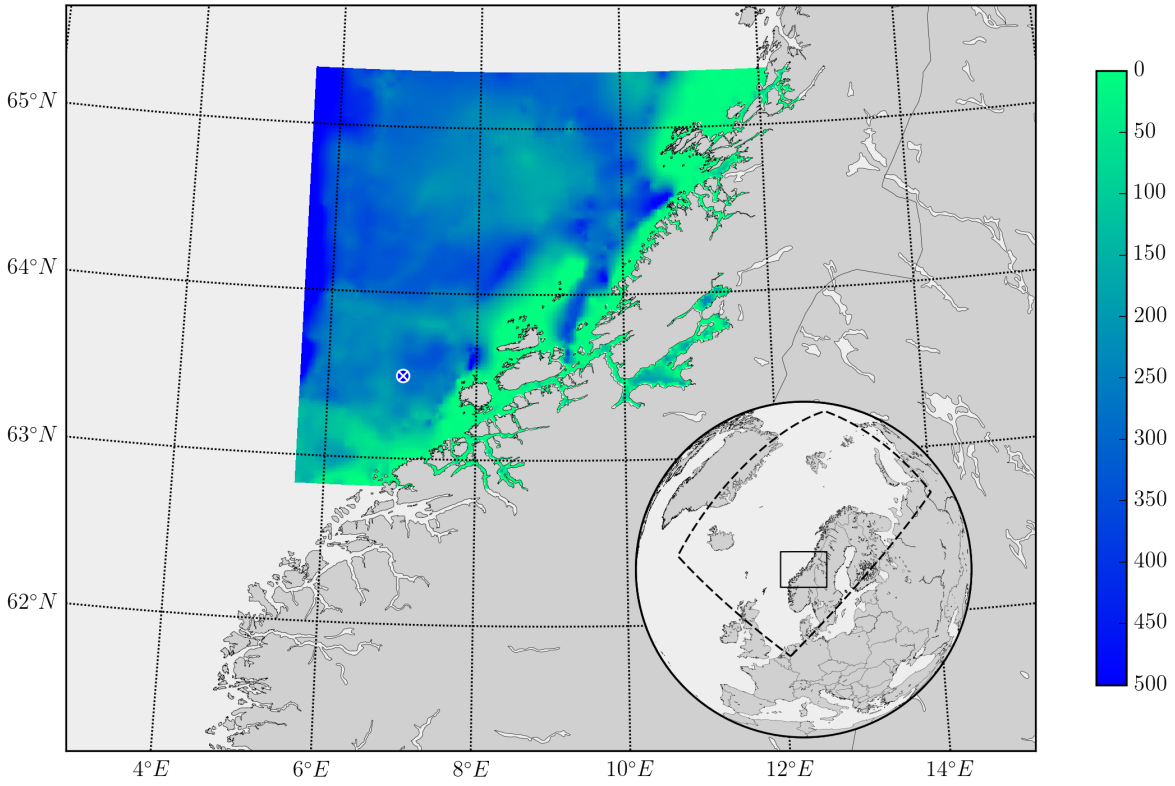

Figure 1 Simulation domain on the coast outside Trondheim, with sea depth (in meters), and location of the release marked. The area of the detailed map is indicated by the rectangle on the inset globe, and the area covered by the current data is given by the dashed outline.

3 Ensemble Simulation Procedure

For each of the three release durations, an ensemble of simulations has been carried out by starting one simulation every hour for a period of four years. The first simulation had a start time of 2009.01.01, 00.00, the next was started at 2009.01.01, 01.00, etc, with the last simulation started at 2012.12.31, 23.00, and running until 2013.01.09, 23.00. This yields a total of 35064 simulations for each of the three release durations.

Each simulation will experience different environmental forcing, leading to different outcomes. Two simulations which are started an hour apart are expected to be very similar, while two simulations started weeks or months apart are less likely to be correlated. By varying the simulation start time in this manner, we sample from the underlying distribution of environmental data, in order to obtain information about the range of possible outcomes of an oil spill at the given location.

The procedure is essentially a form of Monte Carlo simulations, a commonly used scheme in computational science in general, where the defining property is that the input is in some sense “random”. Since we don’t know beforehand what the environmental conditions during a spill will be, we sample repeatedly from historical data, in order to map out the space of possible outcomes. The historical data are random in the sense that we choose only the timing of the release, without letting knowledge of the environmental conditions at that time influence the choice. In principle, a more statistically robust approach would be to generate a truly random and independent realisation of current and wind data for each simulation in the

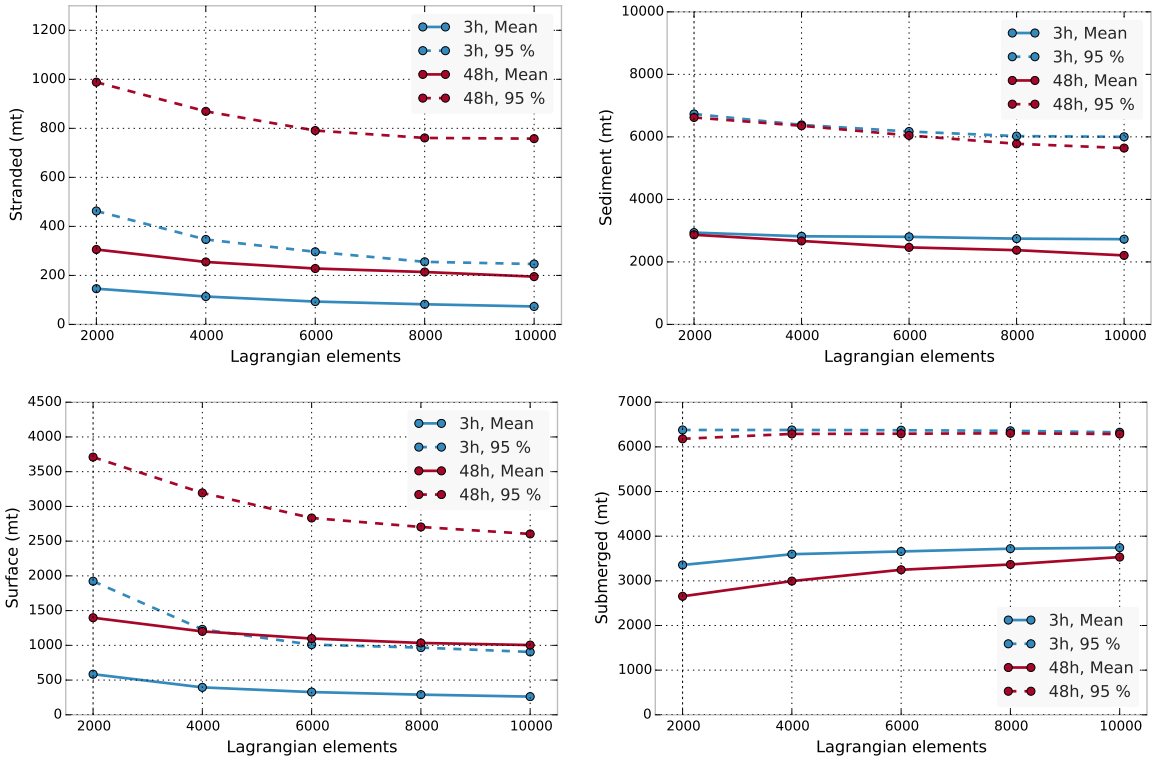

Figure 2 Results of sensitivity analysis, showing mean and 95-percentile for four different compartments of the mass balance, as a function of number of numerical particles (Lagrangian elements) used. The values are calculated from 672 simulations carried out at hourly intervals throughout February 2009, in each case taking the value 10 days after the start of the release. Results shown for 3 and 48 hours release durations.

ensemble. However, to do so would be computationally very demanding, and it is also difficult to generate physically consistent random realisations of environmental data that match the true distribution of these data in an actual location.

While there are several possible strategies for how to sample the historical data, we chose uniform sampling, with a constant time difference of 1 hour between consecutive simulations. By carrying out a very large number of regularly spaced simulations, we are free to later construct smaller subsets, both regularly and in a more random manner. Furthermore, by using a shorter sampling interval than the time step of the environmental data (2 hours), we expect to capture most of the variations possible due to different start times.

The simulation results from the OSCAR model include four dimensional (x, y, z, t) concentration fields giving concentration per component for droplets and dissolved chemicals, as well as three dimensional (x, y, t) grids for oil on the sea surface, on the shore and in the sediments. Additionally, some aggregated quantities are available as time series. These include amounts of evaporated oil, oil on the sea surface, submerged oil, oil on the shore, oil in the sediment and amount of oil which has been biodegraded.

These last six quantities make up what is known as the mass balance, because it gives information about the fraction of the total mass which is found in any given "environmental compartment" (for reasons of linguistic convenience we include biodegraded as a compartment). During the development of a spill, oil can move from one compartment to another. For example, oil on the surface can be mixed down by waves and submerged, submerged oil can

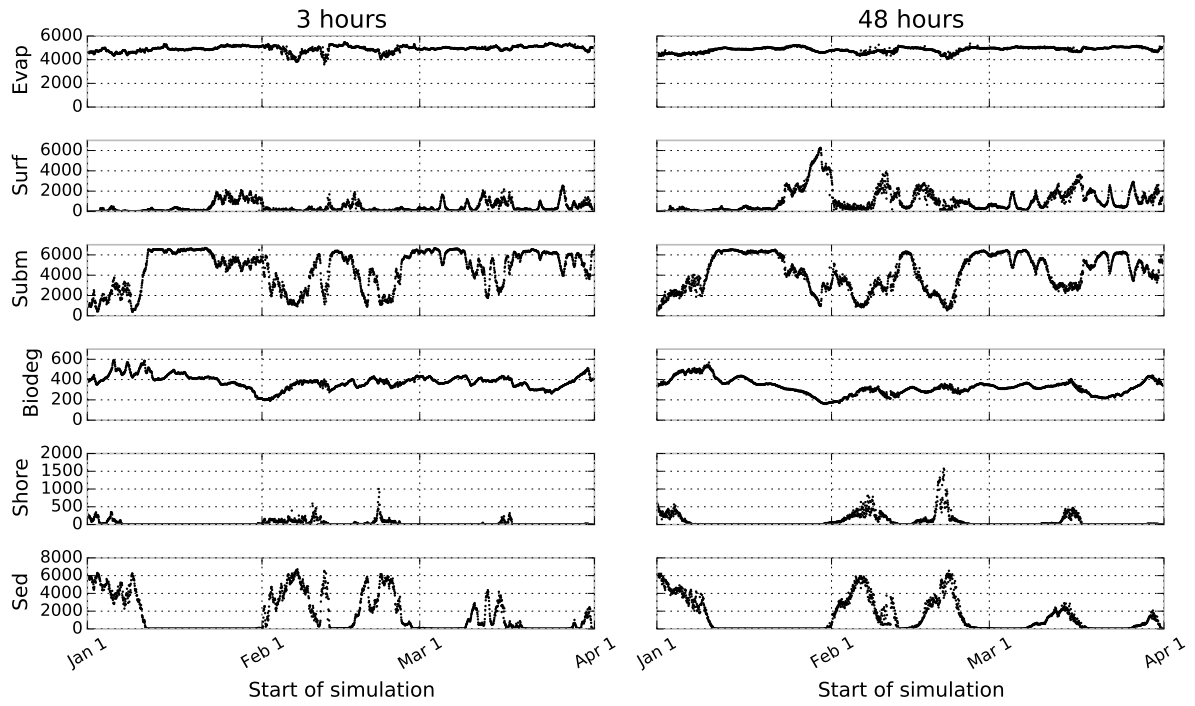

Figure 3 Mass balance (all values in metric tons) for the 3 hour and 48 hour release durations, shown as a function of start time of release, with the values in the figure being taken 10 days after the start of the release. Here shown from January 1 2009 to April 1 2009.

resurface, stranded oil can be washed out to sea, etc. The exception is that oil which has been evaporated or biodegraded is removed from the simulation.

4 Results

A simple sensitivity study was carried out, running hourly simulations throughout the month of February 2009 (chosen somewhat arbitrarily because January 2009 displayed very little stranded oil) with 2000, 4000, 6000, 8000 and 10000 numerical particles (Lagrangian elements). OSCAR employs two types of numerical particles for oil spills, one to represent the dissolved phase (dissolved components from the oil), and one to represent the oil phase (submerged droplets and surface oil). In each case, equal numbers were used, so for the case with 2000 particles, there were 1000 of each type.

For each number of particles, 672 simulations were carried out (other than number of particles, the setup was as shown in Table 1). For each of these simulations, we looked at the mass balance 10 days after the start of the release, and calculated the mean and the 95-percentile for amount of oil on the shoreline, in the sediments, on the surface and submerged (all in metric tons). By 95-percentile, we mean the value which is such that 95% of the results are smaller than this value. The results of the are shown in Figure 2. For the full ensemble of simulations, we chose to run with 10000 numerical particles (5000 dissolved and 5000 oil).

We have not considered spatial distribution in this paper, looking instead at the aggregate properties that make up the mass balance. Thus, we look for example at the total amount of oil on the shoreline, without considering the distribution along the shore. Example time series of such data, showing the six mass balance compartments for the 3 hour and 48 hour release durations, are shown in Figure 3. Each dot in the figure corresponds to one simulated

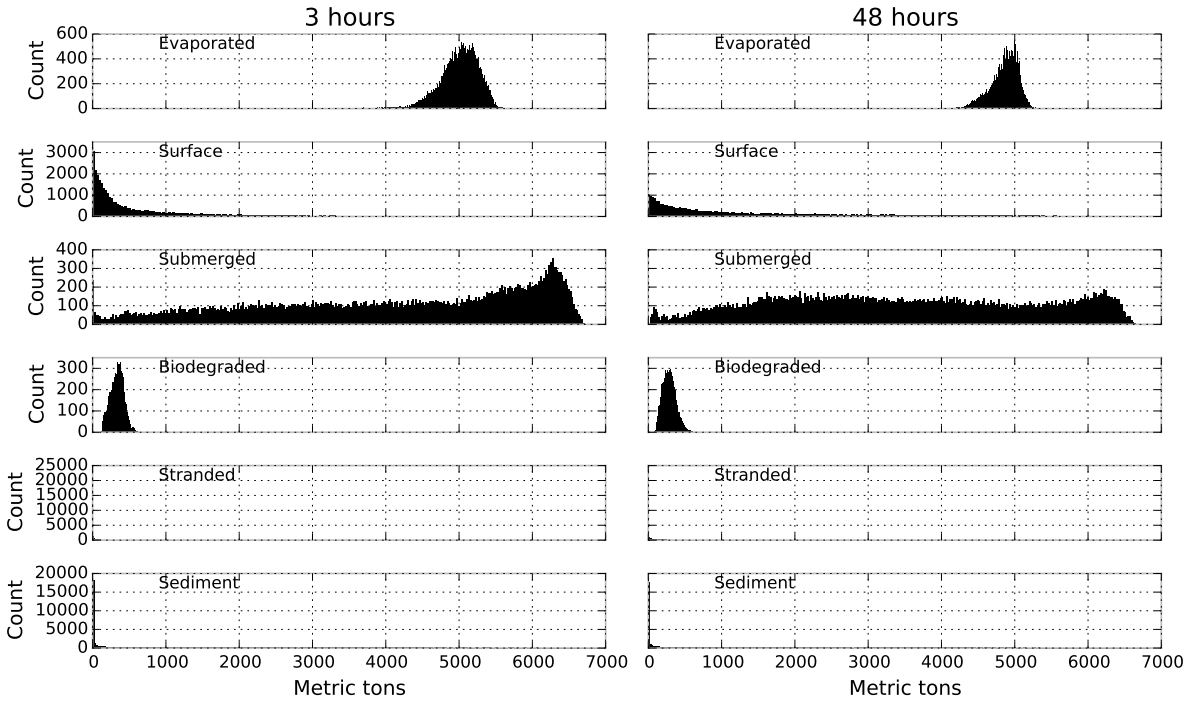

Figure 4 Histogram showing frequency of occurrence of the values of the mass balance for 3 hour and 48 hour release durations, by using the value at the end of each 10-day simulation. This is based on the full ensemble, covering the period 2009 – 2012.

release, with the start time of the release given by the horizontal axis. Hence, the figure does not show the development of a single release. Instead, the value for a given time, t, shows how much oil was found in each of the six compartments, 10 days after the start of the release which began at time t.

We have included all six compartments here because it is interesting to note how they have quite different behaviour. The amount of evaporated oil rarely deviates much from its average value, the amount of biodegraded oil varies by a factor of around 3, whereas the amount of stranded oil is zero or small most of the time, with a handful of samples at very large values. In particular, for the data shown in Figure 3 for amount of oil on the shore, for the 3 hour release duration, the average is 35 mt, the 95-percentile is 185 mt and the worst case is 1001 mt. From these observations, it should be clear that the different compartments obey quite different probability distributions.

In Figure 4, we show the histograms of the distributions of the six compartments, again for the 3 hour and 48 hour release durations, but in this case the figures are based on all the available data, i.e, all 35064 simulations for each duration, covering the years 2009 – 2012. Again, the amount of oil in each compartment at 10 days after the start of the release was used. The results for amount of oil in the sediment, and amount of stranded oil, stand out from the rest by being nearly invisible in this figure. In Figure 5, the histograms for oil in the sediments and stranded oil are shown separately, on a log-log scale.

The goal of an environmental risk analysis is to make statistically robust predictions of, for example, 95-percentile for the amount of oil on the shore. When we would like to estimate the required sample size (meaning number of simulations to include in the ensemble)

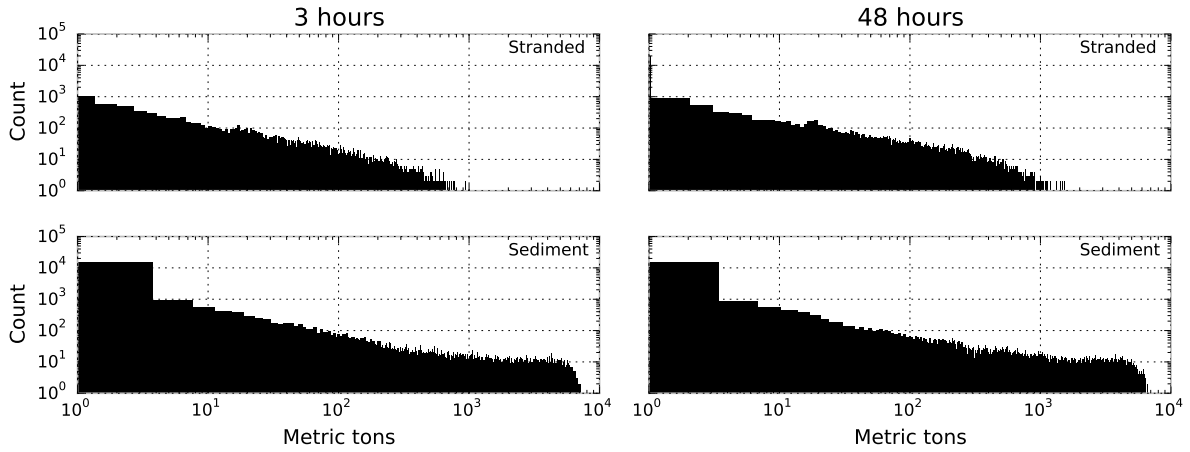

Figure 5 Histogram showing frequency of occurrence of the values of the amount of oil on the shore and in the sediments for 3 hour and 48 hour release duration, by using the value at the end of each 10-day simulation. Shown on log-log scale.

for statistically robust results, it is essential to have an idea of the probability distribution of the variable of interest. For example, from Figures 3 and 4, it seems reasonable to assume that relatively few samples will do if we were interested in estimating the average amount of evaporated oil.

For the amount of oil in the sediments and the amount of stranded oil, however, the situation is quite different. These two seem to follow more fat-tailed distributions, although of course with a cutoff, since we know *a priori* that the largest amount of oil that could end up in either compartment must necessarily be smaller than the total released amount (which in this case is 12000 metric tons). However, even with a cutoff, a property of a fat-tailed distribution is that relatively large values occur much more frequently than in, for example, a Gaussian distribution. As a consequence, a much larger number of samples is required to accurately estimate, e.g., the 95-percentile for amount of stranded oil, as compared to the 95-percentile for amount of evaporated oil.

To further illustrate the differences in distributions, we can compare the average, μ, the standard deviation, σ, the 95-percentile and the 99-percentile for the six compartments. These parameters are presented in Table 2. If we consider for example the amount of biodegraded oil in the 3 hour release, we note that the 95-percentile is about 1.6 standard deviations above the mean, and the 99-percentile is about 2.3 standard deviations above the mean. This is approximately as expected for a Gaussian distribution (1.65σ and 2.33σ respectively). For the amount of oil on the shore, the situation is quite different. We note that the standard deviation for each of the three durations is larger than the mean, even though the variable cannot be negative. Furthermore, the 99-percentile is 3σ, 2.8σ and 2.4σ above the 95-percentile for the 3-hour, 24-hour and 48-hour releases respectively. For a Gaussian distribution, the 99-percentile is expected to be about 0.68σ above the 95-percentile. This serves to illustrate that the normal statistical quantities do not behave as one might expect when dealing with variables which follow fat-tail distributions.

4.1 Correlations

By investigating the autocorrelation in time of our series of simulation results, we get a measure for how "related" two simulations are, as a function of the time, or "lag", between

Table 2 Average, μ, standard deviation, σ, and 95 and 99-percentile for the six compartments of the mass balance. All amounts in metric tons.

	3 hours				24 hours				48 hours			
	μ	σ	95%	99%	μ	σ	95%	99%	μ	σ	95%	99%
Evaporated	4996	279	5377	5466	4902	202	5147	5208	4857	192	5109	5190
Surface	923	1340	3918	6336	1461	1582	4771	6524	1794	1808	5537	6890
Submerged	4066	1866	6407	6566	3683	1803	6314	6496	3486	1754	6242	6466
Biodegraded	335	88.6	474	541	304	87.6	460	530	285	87.9	442	514
Stranded	31.4	90.7	196	466	54.5	145	314	721	67.1	166	400	806
Sediment	949	1651	4953	6144	942	1623	4822	6023	912	1536	4545	5632

them. The autocorrelation, $R(\tau)$, of a discrete signal, X_i, as a function of lag, τ, is given by

$$R(\tau) = \sum_{i=0}^{N-\tau} (X_i - \mu)(X_{i-\tau} - \mu), \tag{1}$$

where N is the length and μ is the average of the signal.

In Figure 7 we show the normalised autocorrelations, i.e., $R(\tau)/R(0)$, of the amount of stranded oil for the three release durations. Here, X_i in Equation (1) is the amount of oil on the shore, at the end of simulation i, when counting from the first simulation in the set. For comparison, we also include autocorrelations for the time series of currents and winds at the release location, shown in Figures 8 and 9.

Looking at Figure 7, we note that the correlation in amount of stranded oil falls off faster for the shorter release durations, meaning that two 48-hour releases separated by an interval of 3 days are more likely to give similar results than two 3-hour releases. This general trend is to be expected from a simple analysis of the problem. Two 3-hour releases, where one starts 24 hours after the other, could in principle end up in completely different locations if, for example, the wind direction had changed in the meantime. For two 48-hour releases with start time separated by 24 hours, however, the environmental conditions will be identical for half the release duration.

For the sake of argument, we might make the assumption that oil released at the same instant, in two different simulations, will always end up in the same location.* We can then argue that the difference in the mass balance of two partially overlapping simulations can only be due to the oil which is released in the two non-overlapping portions of the simulations.

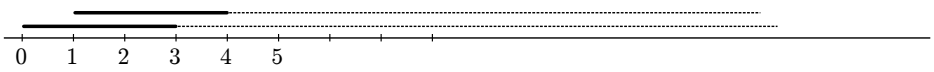

Figure 6 Illustration of two 3-hour releases (thick line) with following simulations (dashed line), started 1 hour apart. The oil released in the overlapping parts of the thick lines will be released into identical environmental conditions.

If we consider again two 3-hour releases (see Figure 6), where one release begins 1 hour after the other, then we would expect the oil released in the two overlapping hours to behave identically in the two cases. Consequently, the greatest absolute difference we could get in, say, the amount of stranded oil, would be if all the oil released in the first hour of the

*In reality, this is not true, since each simulation uses a random walk formulation which means that small differences are to be expected, even under identical conditions, and the movement of the oil will also be affected by the amount of oil, as for example a thick slick on the surface will behave differently from a thin slick.

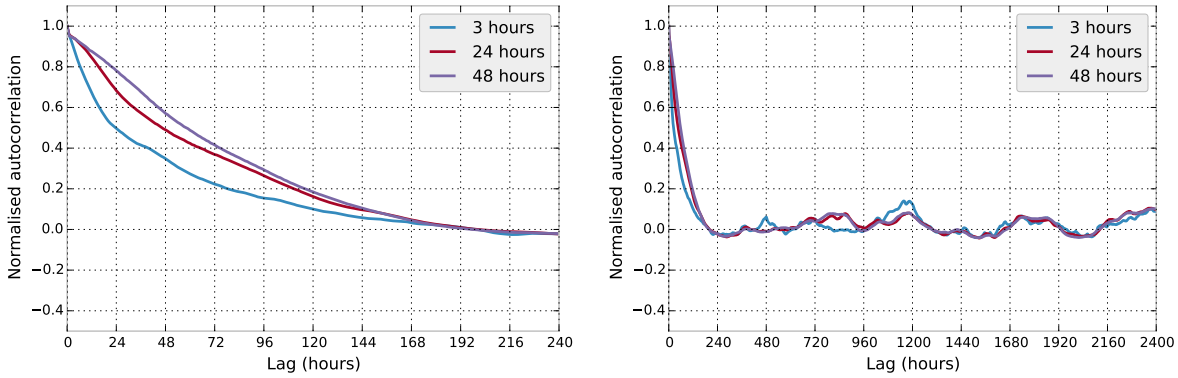

Figure 7 **Normalised autocorrelation of the amount of stranded oil, as a function of lag, shown for the three release durations 3, 24 and 48 hours. Shown for lags from 0 to 10 days (left), and from 0 to 100 days (right).**

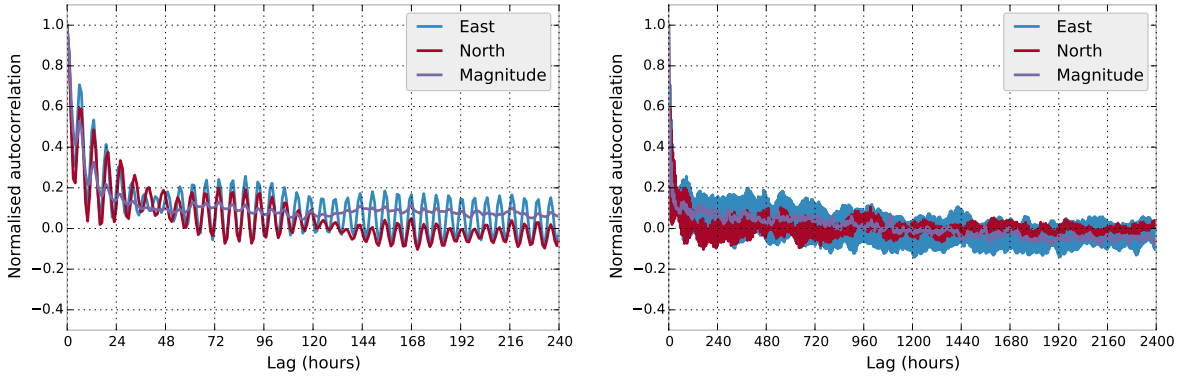

Figure 8 **Normalised autocorrelation of the surface current at the release location. Eastward component, northward component and magnitude shown separately. Shown for lags from 0 to 10 days (left), and from 0 to 100 days (right).**

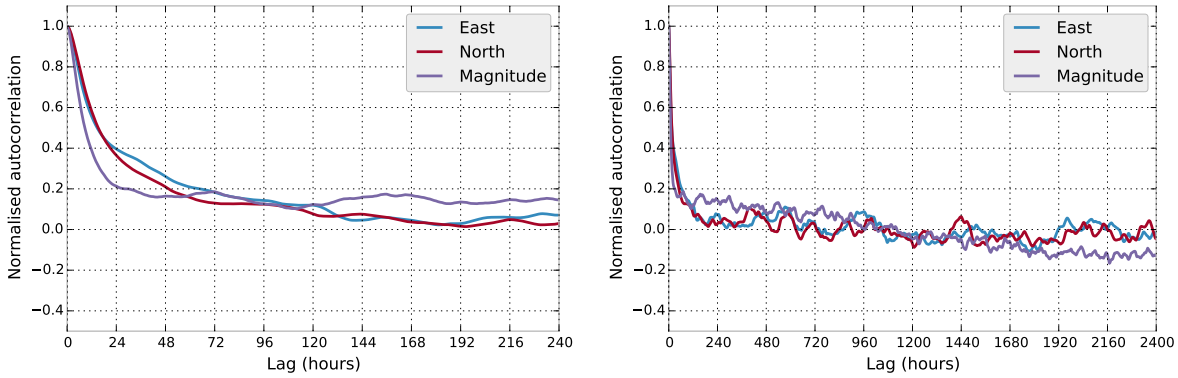

Figure 9 **Normalised autocorrelation of the wind at the release location. Eastward component, northward component and magnitude shown separately. Shown for lags from 0 to 10 days (left), and from 0 to 100 days (right).**

Nordam, T., U. Brönner, and R.L. Daae, Convergence of Ensemble Simulations for Environmental Risk Assessment, Proceedings of the Thirty-ninth AMOP Technical Seminar, Environment and Climate Change Canada, Ottawa, ON, pp. 58-73, 2016.

first simulation ends up on the shore, and none of the oil released in the last hour of the second simulation. Hence, the largest difference in the mass balance of two overlapping simulations is given by the non-overlapping fraction of the total release.

Based on this simplified reasoning, we can estimate how densely we need to sample the available environmental data. Say we are considering a constant 24 hour release, taking place between January 1, 2009 and December 31, 2012, and that we would like to find the worst case scenario, in terms of amount of oil on the shoreline. If we sample the entire four year period, with one simulation every 6 hours, then from the arguments above we can estimate that the "true" worst case will exceed the worst case among our simulations by no more than 12.5% of the total released amount. Since no 24-hour release during that period could differ (in start time) by more than 3 hours from the closest of our simulations, the non-overlapping fraction of the two releases could be no more than 12.5% of the total released amount.

4.2 Sampling Strategies – Uniform Sampling

The idea of ensemble simulations for environmental risk assessment of oil spills is that each simulation is a sample from the space of possible outcomes for that spill. In the following discussion, we use "sampling interval", to mean the amount of time between the start of consecutive simulations in the ensemble. Each individual simulation will of course use a normal model time step and the full resolution of the environmental data, independent of the sampling interval.

In a standard environmental risk assessment, one would normally use a larger sampling interval than once every hour. In order to investigate the effects of the sampling, we can create subsets of our complete set of simulations, and compare statistical parameters, such as the 95-percentile, calculated from subsets constructed in different ways.

For a choice of two hours sampling interval, there are exactly two different ways to construct this subset from our complete set of one-hourly simulations. While starting the first release at 2009.01.01 00.00, and sampling at every even hour, might seem like a natural choice, there is nothing special about this time. We might just as well begin at 2009.01.01 01.00, sampling at every odd hour. In general, for a sampling interval of N_h hours, it is possible to construct N_h different subsets from our complete set of hourly simulations. In constructing the subset, we refer to the start time of the first included simulation, in number of hours after 2009.01.01 00.00, as the offset.

As examples, we now subsample our set of simulations with sampling intervals of $N_h = 120$ hours and $N_h = 240$ hours. Since an offset of 120 hours will give the same subset as an offset of 0 hours, there are 120 possible offsets, and consequently 120 different subsets for $N_h = 120$ hours, and 240 different subsets for $N_h = 240$ hours. In Figure 10, we show the 95-percentile for amount of oil on the shore calculated for each of the 120 (left) and 240 (right) subsets, as a function of offset, for all three release durations. The figure demonstrates that the estimated 95-percentile depends strongly on the particular choice of subset of simulations, especially for the shorter release duration. For the 3-hour release and $N_h = 120$ hours, the largest value (287 mt) is 2.6 times larger than the smallest value (109 mt), while for the 48-hour release, the largest value (498 mt) is only 1.6 times the smallest (320 mt). For $N_h = 240$, corresponding to starting one simulation every 10 days, throughout the four years, the numbers for the 3 hour release are 80 mt (min), 406 mt (max) and 5.1 (max/min), and for the 48 hour release, 237 mt (min), 516 mt (max) and 2.2 (max/min).

We have seen that for a sampling interval of N_h hours, it is possible to create N_h different subsets of our complete set of hourly simulations. Since a smaller sampling interval

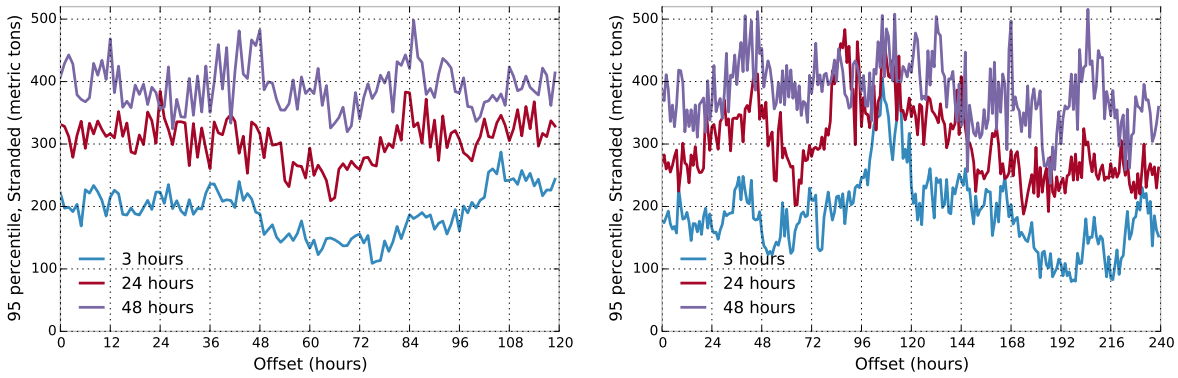

Figure 10 95-percentile for amount of oil on the shoreline (in metric tons), calculated from subsets of the complete sets of simulations, using 120 hours (left) and 240 hours (right) sampling intervals. The offset gives the start time of the first included simulation, in hours after 2009.01.01 00.00.

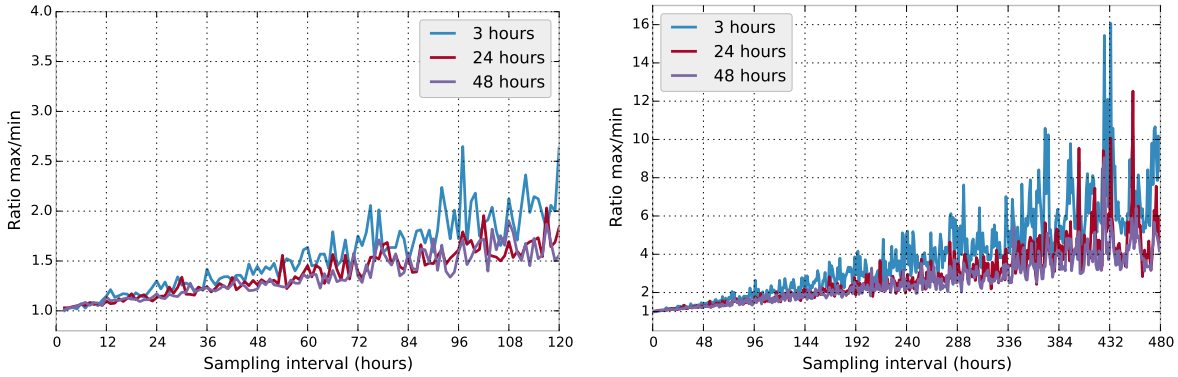

Figure 11 Ratio of highest to lowest 95-percentile for oil on the shore, calculated from the different possible subsets, as a function of sampling interval.

corresponds to more simulations (the number of simulations in a subset with sampling interval N_h hours, is $35064/N_h$), we would expect the results to be more reliable for smaller N_h. An indicator of the reliability of a particular choice of N_h can be found by constructing all the possible subsets, and then calculating N_h different values of the relevant statistic. If the ratio of the highest to the lowest value of this statistic is large, the result is strongly dependent on which of the N_h different offsets you have chosen. Since the offset is a completely arbitrary choice, it should ideally not affect the results.

In Figure 11, we show the ratio of the highest to the lowest value of the 95-percentile for amount of oil on the shore, calculated from each of N_h subsets of our complete set of simulations, as a function of N_h. An analysis such as this can be used to inform the choice of sampling interval. For example, these results indicate that for the 3-hour release we have studied here, if one chooses a 96 hour sampling interval, one might expect an error of about a factor 2 - 2.5 in the calculated 95-percentile for oil on the shore, simply due to the arbitrary choice of offset. For the 48-hour release, on the other hand, a sampling interval of 96 hours gives a possible variation in the predicted 95-percentile of only a factor 1.5.

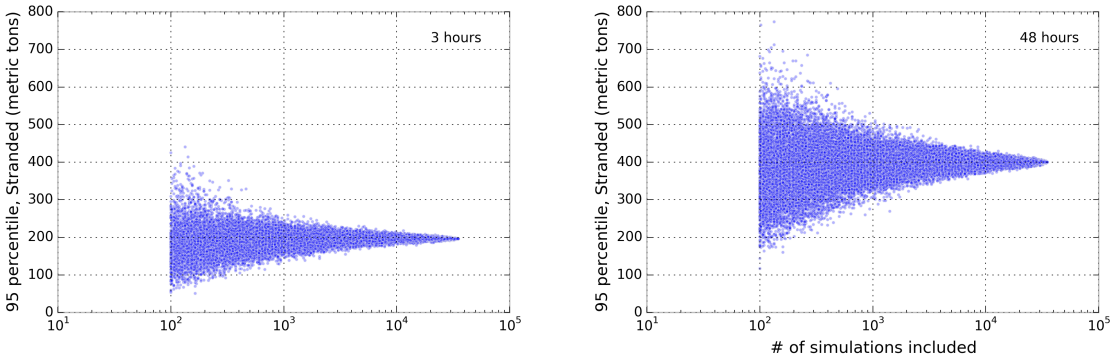

Figure 12 95-percentile for amount of oil on the shoreline (in metric tons), calculated from 200 randomly generated subsets of a given size, at 200 different subset sizes ranging from 100 to 35064. To the left, for a 3 hour release, to the right, for a 48 hour release.

4.3 Sampling Strategies – Random Sampling

Another sampling strategy consists of carrying out an ensemble of simulations with random release start times. By generating random subsets of our complete set of simulations, we can also investigate how this strategy compares to uniform sampling, although since we are limited to choosing from a base set of simulations with uniform spacing, it will not be truly random sampling.

A subset of size N_s, of our complete set of $N = 35064$ simulations, can be selected in a large number of different ways ($N!/N_s!(N - N_s)!$). In Figure 12 we show the 95-percentile for amount of oil on the shore, calculated from 200 randomly generated subsets for each of 200 subset sizes, starting at $N_s = 100$ and going up to $N_s = 35064$, i.e., including all simulations from the complete set.

As the number of included samples increases, the 95-percentiles calculated from the subsets converge towards the 95-percentile in the complete set of simulations. While this does not guarantee that the 95-percentile of the complete set of simulations is identical to the "true" 95-percentile, it does indicate that the number of included simulations is sufficiently large that a few outliers will not significantly alter the calculated percentile.

To illustrate the point that apparent convergence does not guarantee convergence to the "true" result, we have split the set of simulations into four, giving one set of 8760 simulations for each of the years 2009 – 2012 (strictly 8784 for 2012, which was a leap year). We then generated the same convergence plot as shown in Figure 12, but now for each year separately, using 100 values of N_s, and 100 realisations for each value. The results are shown in Figure 13. As can be seen from these figures, the 95-percentile for each individual year appears to converge smoothly, but they converge towards different values. In particular, the 95-percentiles of amount of oil on the shore for the simulations carried out during 2010 and 2012 are 2-3 times higher than for 2011, for both the 3-hour and the 48-hour releases. This demonstrates the role of inter-annual variability in the environmental data, and highlights the need for using sufficiently long data series

5 Discussion and Conclusions

In deciding upon a strategy to carry out an ensemble of simulations for environmental risk assessment, there are a number of important questions to ask. Among them are "How long a time series of environmental data to use?" and "How to sample the environmental data?".

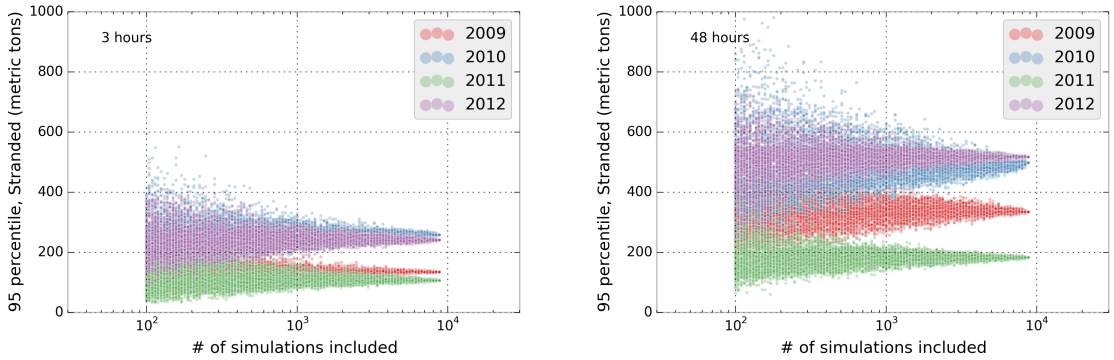

Figure 13 95-percentile for amount of oil on the shoreline (in metric tons), calculated from 100 randomly generated subsets of a given size, at 100 different subset sizes ranging from 100 to 8760, calculated separately for each of the four years 2009 – 2012. To the left, for a 3 hour release, to the right, for a 48 hour release.

We have argued that the sampling interval (meaning time interval between the start of consecutive simulations in the ensemble) depends on the release duration, with shorter durations requiring a shorter sampling interval. This can be inferred from the auto-correlations, shown in Figure 7, as well as from physical considerations of the behaviour of two partially overlapping releases. The fact that a shorter release requires denser sampling to achieve the same relative accuracy is illustrated by Figure 11, which shows that the shorter releases have a wider range of predicted 95-percentiles, for a given sampling interval. We also argue that the amount of stranded oil, and the amount of oil in the sediments, follow different probability distributions from, for example, the amount of biodegraded oil, and that as a consequence, a larger number of samples is required to accurately estimate the amount of oil in these compartments.

We have also demonstrated that even though the complete set of simulations for the years 2009 – 2012 appears to converge smoothly towards a well-defined 95-percentile when a large number of simulations is considered, the four years considered separately converge towards different values, as illustrated in Figure 13. The large difference seen between the individual years indicate that we should probably have used more than four years, if we were carrying out an actual environmental risk assessment for this case. For Northern Europe, inter-annual differences could (among other things) be due to the North Atlantic Oscillation, which influences trends in wind speeds and directions over a timescale of years. A follow-up study to this work is planned, where we will be able to use 10-20 years of high resolution environmental data. Additionally, an HPC cluster currently being installed at SINTEF will provide us with the computational resources to also look at the effects of varying the release position.

6 Acknowledgements

The work described in this paper was carried out in an internally funded SINTEF project. The authors would like to thank their colleagues CJ Beegle-Krause, Emlyn Davies, Raymond Nepstad, Mark Reed, Petter Rønningen, Jørgen Skancke, Christos Stefanakos, and Josip Zoric for many an interesting discussion in the SINTEF CoffeeLab.

7 References

Barker, C., “A statistical outlook for the Deepwater Horizon oil spill”, Monitoring and Modeling the Deepwater Horizon Oil Spill: A Record-Breaking Enterprise, pp. 237–244, 2011.

Barker, C. H. and W. P. Healy, “Statistical analysis of oil spill response options: A noaa-us navy joint project”, International Oil Spill Conference Proceedings, American Petroleum Institute, 883–890, Tampa, FL, 2001.

Brandvik, P. J., Ø. Johansen, F. Leirvik, and U. Farooq, “Droplet Breakup in Subsurface Oil Releases – Part 1: Experimental Study of Droplet Breakup and Effectiveness of Dispersant Injection”, Marine Pollution Bulletin, 73:319-326, 2013

Dee, D., S. Uppala, A. Simmons, P. Berrisford, P. Poli, S. Kobayashi, U. Andrae, M. Balmaseda, G. Balsamo, P. Bauer, et al., “The ERA-Interim Reanalysis: Configuration and Performance of the Data Assimilation System”. Quarterly Journal of the Royal Meteorological Society, 137:553–597, 2011.

Egbert, G. D., A. F. Bennett, and M. G. Foreman, “TOPEX/POSEIDON tides estimated using a global inverse model”, Journal of Geophysical Research: Oceans (1978–2012), 99:24821–24852, 1994.

Guillen, G., G. Rainey, and M. Morin, “A simple rapid approach using coupled multivariate statistical methods, GIS and trajectory models to delineate areas of common oil spill risk”, Journal of marine systems, 45:221–235, 2004.

Johansen, Ø., M. Reed, and N. R. Bodsberg, “Natural dispersion revisited”, Marine Pollution Bulletin, 93:20-26, 2015

Johansen, Ø., P. J. Brandvik, and U. Farooq, “Droplet Breakup in Subsurface Oil Releases – Part 2: Predictions of Droplet Size Distributions With and Without Injection of Chemical Dispersants”, Marine Pollution Bulletin, 73:327-335, 2013

Johansen, Ø., H. Rye, and C. Cooper, “DeepSpill – Field Study of a Simulated Oil and Gas Blowout in Deep Water”, Spill Science & Technology Bulletin, 8:433-443, 2003

Price, J. F., R. A. Weller, and R. Pinkel, “Diurnal Cycling: Observations and Models of the Upper Ocean Response to Diurnal Heating, Cooling, and Wind Mixing”, Journal of Geophysical Research: Oceans (1978–2012), 91:8411–8427, 1986.

Price, J. M., W. R. Johnson, C. F. Marshall, Z.-G. Ji, and G. B. Rainey. “Overview of the oil spill risk analysis (OSRA) model for environmental impact assessment”, Spill Science & Technology Bulletin, 8:529–533, 2003.

Price, J. M., W. R. Johnson, Z.-G. Ji, C. F. Marshall, and G. B. Rainey, “Sensitivity testing for improved efficiency of a statistical oil-spill risk analysis model”, Environmental Modelling & Software, 19:671–679, 2004.

Reed, M., P. Daling, O. Brakstad, I. Singsaas, L. Faksness, B. Hetland, and N. Ekrol, “Oscar 2000: A multi-component 3-dimensional oil spill contingency and response model”, Proceedings of the 23rd AMOP Technical Seminar, Environment Canada, Ottawa, ON, 1:663–680, 2000.

Slagstad,D. and T. A. McClimans, “Modeling the Ecosystem Dynamics of the Barents Sea Including the Marginal Ice Zone: I. Physical and Chemical Oceanography”, Journalof Marine Systems, 58:1-18, 2005.

Smagorinsky, J., “General Circulation Experiments With the Primitive Equations: I. The Basic Experiment”, Monthly Weather Review, 91:99–164, 1963.

Spaulding M. L., C. Galagan, L. McStay, T. Isaji, D. Stuebe, and R. Green, “Use of Ensemble Based Methods for Oil Spill Risk Assessment Investigations”, Proceedings of the 37th AMOP Technical Seminar, Canmore, AB, 2014

World Meteorological Organization, “Guidelines on ensemble prediction systems and forecasting”, Technical report, WMO, Geneva, Switzerland, 2012.