Ocean Science Discussions

Ocean Sci. Discuss., author comment AC3
https://doi.org/10.5194/os-2021-83-AC3, 2021

Reply on RC1

Mikael L. A. Kaandorp et al.

Author comment on "Using machine learning and beach cleanup data to explain litter quantities along the Dutch North Sea coast" by Mikael L. A. Kaandorp et al., Ocean Sci. Discuss., https://doi.org/10.5194/os-2021-83-AC3, 2021

Summary

The manuscript uses machine learning techniques to formulate a regression model to investigate the influence of a range of hydrodynamic, atmospheric and coastline geometry factors obtained from publically available data and re-analysis products, as well as particle tracking model runs based on publically available hydrodynamics model output, on the density distribution of (plastic) beach litter on the Dutch coast. The regression model is 'trained' using several years of data from beach cleaning surveys. The main explanatory factors identified are negative correlations with tidal elevation variability and maximum tidal amplitude, and a correlation with coastline orientation relative to residual currents. The particle tracking model runs did not contribute significantly to the predictive capability of the regression model. A trimmed-down version of the regression model, using only the most important contributing factors, was used to generate a 5-year average litter concentration distribution along the Dutch coast.

General remarks

This is a very well-written manuscript, that I have read with interest. It mobilises a variety of different data, and uses sophisticated statistical techniques to analyse them together. It addresses a clear need to improve understanding and predictive capability of beaching plastics in the marine environment. Having said that, the manuscript is not completely convincing, and leaves me wondering about a few things:

We would like to thank the reviewer for this positive feedback on our manuscript, and the thorough remarks and comments which we will discuss below.

- **By which mechanism do tidal variability and amplitude control the beach litter distribution? Do higher variability and amplitude promote removal of previously beached material and in that way define a lower equilibrium concentration? Or does it simply spread the litter more in the cross-shore direction making it harder to detect in the beach surveys? How might this be**

illustrated? Or is the preference of the regression model for tidal variability and amplitude an artifact of the method (see also next points)?

We have addressed the different effects that tides might have on litter concentrations in the main text (line 314 track changes):

In general, a higher tidal maximum and variability lead to less litter measured on the coastline (see the Appendix B5 for further details). A higher tide during or preceding the cleanup could re-suspend some of the litter from the beach. Furthermore, a higher tide encountered during the cleanup stage reduces the beach width that can be sampled. Perhaps a stronger variability in the tidal height leads to less persistent high strandlines where the highest litter concentrations are normally found (Heo et al., 2013). It has been shown in numerical studies that residual tidal currents can lead to a net transport of both suspended and floating matter (Gräwe et al., 2014; Børve et al., 2021; Schulz and Umlauf, 2016). While the regression model indicates that tides play an important role, it is difficult to separate the causal relations between all these different effects and the litter quantities found on beaches. To quantify this in more detail, further experimental and numerical studies are required.

We have added a paragraph in the conclusions and recommendations discussing that further experimental and numerical studies are needed to investigate the effect of tides, since we do not know the exact causal effect (line 475 track changes):

Future studies could further investigate the causal relations between the variables seen as important predictors by the regression model and the litter concentrations found on beaches. This is especially the case for tides, which constitute the two most important features in the regression model (see Figure 5). Experimental studies could further determine whether lower litter concentrations at locations with higher tidal variability are mainly caused by litter re-suspending back into the sea, or for example due to the fact that less area of the beach is sampled during high tide. It should additionally be investigated how these effects compare to the role of (residual) tidal currents, as it has been shown that this can play an important role in transporting suspended matter towards the shore (Schulz and Umlauf, 2016). Experimental investigations can be done in combination with numerical studies of the nearshore marine environment, to capture the interactions between processes such as tides, waves, and particle sizes (Alsina et al., 2020).

- **A regression model cannot be used to prove causality. Indeed, the authors acknowledge that many of the factors investigated are not independent, as well as admitting that there are many regression factors compared with the number of available data points. I think that the authors can do more to make their case:**

a) they can 'cripple' the regression model by removing (some of) the most important factors (rather than those that do not matter much) and show that the resulting model has significantly reduced predictive capability (as opposed to shifting predictive power to a different factor);

b) they state that the particle tracking model does not add to the predictive capability as it includes most of the factors from others sources incorporated in the regression model. So they can treat the particle tracking model as an independent (causal!) experimentation tool, and validate its results against the field data as well as cross-comparing with the results of the regression model, and subsequently disable the most important processes suggested by the regression model to directly illustrate their influence on the results.

The reviewer is correct that this is indeed a difficult aspect of using data-driven machine learning techniques to make predictions, as opposed to traditional physical models, where it is easier to draw conclusions on causal relations between observations and certain model parameters or processes.

We have included a figure in the supplementary information now (Figure B5), where we checked what happens when leaving out certain important features from the regression model. We have also added (line 545 track changes):

In Figure B5, we analyse the effect of leaving out certain feature categories on the model performance. The random forest can create a highly non-linear map between the features and corresponding response. It is therefore possible that when using a large set of features and leaving out one important explanatory variable, it will use a combination of the remaining features to still obtain a good fit. We therefore only use the top 10 features in this analysis, and exclude the Lagrangian model variables, as these implicitly contain information on the other features. As can be seen, leaving out a certain category of features reduces the model performance. This can especially be observed when leaving out all features regarding tides, and the two features regarding coastal properties (l_{coast} and $n_{grid} \cdot n$). The mean Pearson correlation coefficient decreases and the variance of the model performance increases.

Issue b) would be very interesting to check, but this is out-of-scope for this project. One of the points of this study was to investigate what kind of processes might play an important role near the coastline regarding the beaching of litter. Once we have a better understanding of this, we can include these processes to model the beaching of litter more accurately. As we discuss in the main text (line 148 track changes), most contemporary particle tracking simulations use e-folding time scales to model how quickly litter might end up on the coastlines. These time scales are assumed to be constant for now, which is of course a simplification. Since these particle tracking simulations do not yet contain most of the relevant parameters in the beaching parameterization, we can also not disable them and cross-compare it with the regression model. But this is definitely something which should be implemented and investigated in the future.

We are currently further investigating whether we can indeed combine our particle tracking models with machine learning models, and cross-compare the results, see for example doi.org/10.5194/egusphere-egu21-274.

- **Have all potential causal factors been included in the analysis? The particle tracking results suggested that rivers may be the dominant source of beach litter. One factor I'm missing (and that may be as important or more important than tidal variability) is the proximity of the nearest riverine source (in the upstream direction). Should this not be added as a regression factor? It's possible that in the current setup the influence of a factor like this is attributed to the tidal factors which have a spatial gradient in the same direction.**

This is a very good suggestion. We have re-run our model, with salinity as a proxy for the distance to the closest river. By doing this, a low salinity indicates that there is a river close-by in the upstream direction.

We have also re-run our Lagrangian simulations, per the comments we received on a related paper in Geophysical Research Letters (for the preprint see DOI: 10.1002/essoar.10508992.1). In this paper the reviewers noted correctly that the dataset used for the currents already contains the net effect on the ocean flow induced by tides. We therefore removed the additional tidal forcing in our Lagrangian model simulations. This is adjusted in the main text (line 129 track changes):

We do not add additional tidal forcing to the Lagrangian model (Sterl et al., 2020) since the net effect of tides is already included in the ocean surface current data set (Global Monitoring and Forecasting Center, 2021).

One of the first things to note, is that more importance is given to the litter fluxes calculated from the Lagrangian model runs. This is good news, as it indicates the Lagrangian model runs have greater predictive power compared to what we saw previously. We now also see that the model attributes more of the plastic pollution to originate from fishing activity and coastlines, which is exactly in line with what we found in our related paper (10.1002/essoar.10508992.1): about 40% of the plastic litter in the Netherlands is observed to originate from the fishing industry. More importance is attributed now to litter fluxes from fisheries and coastlines compared to rivers. Furthermore, proximity to rivers, using the salinity variables we introduced, is not seen as an important predictor by our model.

- **Is the regression model as good as it seems? I would like to see direct comparisons of predicted annual distributions with the observed annual spatial distributions in Appendix A. I'm aware that these data were used for training so it's not a validation, but it would illustrate to what extent the regression model captures the temporal and spatial variations, which are not represented in Figure 4.**

We have added figures to appendix A comparing the litter predictions with the litter observations per year per location. We plot the test data points here for each of the outer loops in the 5-fold cross validation process, so technically these points are not contained in the training data. We will use the illustration below (example year 2017) to discuss some of the things that the model gets right, and some of the things that the model is missing.

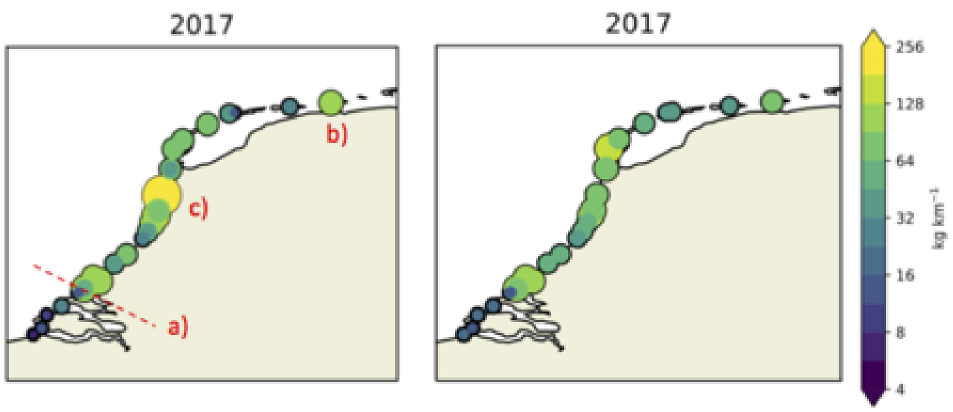

Fig R1: Comparison 2017, observations (left) and modelled results (right), taken from Fig A2 in the manuscript.

Firstly, the lower litter concentrations in the south are reproduced well by the model. Furthermore, the model accurately captures the boundary at which litter concentrations strongly increase, marked by a) in Figure R1. North of this boundary, litter concentrations are generally high, which is reproduced well by the model for all years. In the observations the litter concentrations are often slightly higher on the eastern most island (Schiermonnikoog), marked by b) in Figure R1, which is also reproduced quite well by the model for the various years. One thing that the model misses are outliers, of which one example is marked by c) in Figure R1. Exceptionally high litter concentrations can be caused by all kinds of factors, such as very local 'point sources' of litter which are not captured by models, or subgrid-scale processes. We also discussed this subgrid-scale variability using the variogram analysis. Looking at Fig 4 in the manuscript, it can already be seen that the model misses some variability: some outliers can be observed of very low observed litter concentrations (the points in the upper left corner of the scatter plot outside of the 2σ lines), and very high observed litter concentrations (the points in the lower right corner of the scatter plot outside of the 2σ lines).

We have added subsequent discussion to the main text (line 302 track changes):

It can be seen that there are two kinds of outliers in Figure 4: low observed litter concentrations not captured by the model (points in the upper left corner of the scatter plot), and high observed litter concentrations not captured by the model (points in the lower right corner of the scatter plot). This can be explained by the fact that the model is not able to capture all variability contained in the observations. As the hydrodynamic and wind data in the model have a limited resolution, subgrid-scale effects are missing (see Section 4.2). Furthermore, local point sources of litter (both spatially and temporally, e.g. shipping container accidents) are not captured by the model.

I would like to challenge the authors to address these points, which would significantly improve the manuscript, and hence recommend major revisions.

Specific comments

l 10-12: recommendations for removal. You currently don't really give any, please remove this sentence from the abstract or work out clear recommendations and state them specifically here.

We have added a specific recommendation to the abstract.

l 77 Some sites were sampled multiple times per year. Please describe how these were treated in the analysis. Would this bias the results and how?

In cases where multiple data points are plotted per year, different stretches of beach were cleaned up. For example, once the northern stretch of beach, once the southern stretch. However, these individual stretches are too small to be plotted on the map, which is why some stages contain multiple points per year. We have added this clarification in the caption as well:

For stages with multiple data points per year, different stretches of beach were cleaned (e.g. once the northern side, once the southern side).

Table 1. daily mean currents. This would aliase in a tidal component. What would be the influence on the results? Also, are the bathymetries used to produce the various products used consistent?

The North West Shelf Reanalysis product contains the net effect of 15 tidal constituents. Therefore the net tidal effect is captured in the Lagrangian simulations. These Lagrangian simulations are meant to capture the large-scale particle transport at long time scales. In the statistical analysis we can then use the FES2014 data to calculate the tidal components at high frequency to assess their influence. The Waves reanalysis and ERA reanalysis products use the same bathymetry data (ETOPO2), the North West Shelf reanalysis product uses the GEBCO bathymetry data, the FES2014 data uses a composite, so these might differ slightly. However, what we consider to be more important, is that the atmospheric forcings are consistent (all obtained from the ERA5 data), and that all data are assimilated. We have added text for clarification (line 111 track changes):

High frequency tidal forcing has been used to produce the ocean current data, but output is only provided daily. To capture the effects of tides on a high temporal resolution, FES2014 data are used. Tidal currents (U_{tides}) and heights (h_{tide}) are calculated, taking the M_2, S_2, K_1, and O_1 constituents into account (Sterl et al., 2020), as well as the M_4 and M_6 components which have been shown to play an important role in transport of suspended particles in the North Sea (Grawe et al, 2014).

Figure 2 / rivers: why has only a subset of rivers been included? What has driven the decision to include certain rivers but not others? Please include a list of rivers used in an appendix?

Only rivers which transport more than 0.2 tonnes of plastic litter into the ocean according to the Lebreton dataset are plotted in the figure. We have added this as clarification to the caption:

While all rivers from Lebreton et al. (2017) are included in our analysis, only rivers predicted to transport more than 0.2 tonnes of plastic litter into the ocean are plotted here.

Please note that we did not produce these riverine data, which are provided by Lebreton et al. at doi:10.6084/m9.figshare.4725541

equation 1. In reality, I would expect tau_beach=F(x,y,z,t)? Please discuss why a constant was chosen?

This is a good point. One aim of this study is to get a better understanding under what conditions we can expect more litter to beach, since at the moment this is not well understood. This means we also do not yet know how this beaching time scale can be best parameterized. Only after getting a better understanding of the important beaching processes, we can attempt to improve this parameterization. So, in other words, moving from a constant to a more complicated function could be follow-up work to this manuscript. We have added a clarification to the text (line 148 track changes):

While in reality τ_{beach} might vary significantly both in space and time, it is unknown how this can be best parameterized (Onink et al., 2021). We use the Lagrangian model simulations to capture the large-scale transport of litter, and allow the regression model to pick the most appropriate value for τ_{beach} later on.

Figure 6 and associated text: it would be helpful to have an indication what the principle components may dominantly represent if that's at all possible?

We have investigated the correlations between the principal components and the features, and added a discussion of this to the text (line 384 track changes):

The first principal component shows the highest absolute correlation (Pearson R: 0.45) with long-term tidal variability (with a lead time of 30 days). The second principal component shows the highest absolute correlation (Pearson R: -0.58) with the nearby coastal length (within a radius of 50km). As the measurements taken between 52-53°N are clustered quite closely together, this indicates that conditions regarding tides and coastline geometry are relatively similar for these locations.

l 67 month: which month of the year?

We have added a clarification to the text:

During this tour, every year in August, the entire Dutch North Sea Coast is cleaned up by volunteers.

l 82 August: why one month? Why this month?

See comment above. This is the month that the tour is always organized.

Technical corrections

l 5. what: which

adjusted

l 6 remove might

adjusted

l 6 variability: of what? types/size/volume/??

added: beach litter concentrations

l 10 what: which

adjusted

l 15 increase: release?

adjusted

l 15 need of: need for

adjusted

l 21 other sinks: such as?

Changed to: *In addition, the plastic concentrations found on beaches are generally higher compared to other environmental compartments such as the surface water or the seafloor…*

l 23 …plastic items by removal, …

adjusted

l 25 …reduction of new plastic waste…

Awareness can also lead to more people picking up old plastic waste, so we leave it as is

l 75 weighing devices: scales

adjusted

l 80 mean currents: surface? depth-averaged?

added: mean surface currents

Figure 1. Please include a scale vector for the currents.

A scale vector was added to the figure

Figure 1 caption. Please include references to the data sources. Add a period to the end of the caption.

Data source references were added

l 107 there's a space between models and the period ending the sentence.

adjusted

l 108 the daily-mean ocean-surface currents

l 109-111 include references to the data sources?

Regarding the two comments above: the temporal resolution and references are stated in the table to make the text more readable

l 134 radius of 50km: from where?

Added '...*from the coastline*...'

Many of the figures: please re-consider the choices of colours to facilitate colour-blind people.

We thank the reviewer for this comment, we weren't aware that the standard quantitative colormap in python is not colorblind friendly. We use a colorblind friendly version now. Also we use greyscale for the rivers in Figure 2 now.

Figure 2: please include axes.

We have added longitude/latitude axes

l 187 what: which

adjusted

l 188: distance: between what?

Added separation distance between the different cleanup locations...

l 345 less: fewer

adjusted

l 403 insights into which processes may be causing

adjusted

l 404 and which length scales should be

adjusted