Ocean Modelling 29 (2009) 269–276

Contents lists available at ScienceDirect

Ocean Modelling

journal homepage: www.elsevier.com/locate/ocemod

Lagrangian validation of numerical drifter trajectories using drifting buoys: Application to the Agulhas system

E. van Sebille [a,*], P.J. van Leeuwen [a,1], A. Biastoch [b], C.N. Barron [c], W.P.M. de Ruijter [a]

[a] *Institute for Marine and Atmospheric Research Utrecht, Utrecht University, The Netherlands*
[b] *Leibniz Institute of Marine Sciences, Kiel, Germany*
[c] *Naval Research Laboratory, Stennis Space Center, Mississippi, USA*

ARTICLE INFO

Article history:
Received 20 March 2009
Received in revised form 19 May 2009
Accepted 20 May 2009
Available online 28 May 2009

Keywords:
Numerical Lagrangian drifters
Model validation
Kolmogorov–Smirnov statistics
Drifting buoys

ABSTRACT

The skill of numerical Lagrangian drifter trajectories in three numerical models is assessed by comparing these numerically obtained paths to the trajectories of drifting buoys in the real ocean. The skill assessment is performed using the two-sample Kolmogorov–Smirnov statistical test. To demonstrate the assessment procedure, it is applied to three different models of the Agulhas region. The test can either be performed using crossing positions of one-dimensional sections in order to test model performance in specific locations, or using the total two-dimensional data set of trajectories. The test yields four quantities: a binary decision of model skill, a confidence level which can be used as a measure of goodness-of-fit of the model, a test statistic which can be used to determine the sensitivity of the confidence level, and cumulative distribution functions that aid in the qualitative analysis. The ordering of models by their confidence levels is the same as the ordering based on the qualitative analysis, which suggests that the method is suited for model validation. Only one of the three models, a 1/10° two-way nested regional ocean model, might have skill in the Agulhas region. The other two models, a 1/2° global model and a 1/8° assimilative model, might have skill only on some sections in the region.

1. Introduction

Assessing the skill of ocean models is an important step before the data produced by such a model can be analyzed and interpreted. Special projects have been set up to facilitate the comparison of different ocean models within a fixed framework (e.g. the Coordinated Ocean-ice Reference Experiments (CORE), Griffies et al. (2009)). One of the problems of such skill assessment is that the observations to which the model should be verified are scarce in space and time. The skill assessment is therefore, often limited to a subset of the state vector.

Historically, verification is predominantly qualitative, where one or more specific model variables are compared to observations of these variables. The advantage of this qualitative method is that it introduces the expertise of the modeler in selecting fields and regions that are more important than others. However, the qualitative method also introduces subjectiveness into the skill assessment procedure.

There are objective methods to assess the model skill. Hetland (2006) introduced a way to calculate the improvement of a model with respect to some climatology. Using statistics on the complete model domain, however, has the disadvantage that dynamically relevant regions (such as the western boundary currents) are treated similar to dynamically less important regions. This is a relevant problem especially when the subsequent data analysis is done using numerical Lagrangian floats, tracers that are advected with the flow. These floats often cluster in some regions of the model domain and only the model skill in these regions is relevant for the aptitude of the float data. Ideally, these regions should, therefore, have more weight in the skill assessment. A way to accomplish this focus on dynamically relevant regions is to base the skill assessment on the float trajectories themselves.

The assumption behind trajectory verification is that only skillful models produce trajectories with similar properties as drifting buoys. Therefore, a high skill in float trajectories implies that the underlying model is highly skilled. Here, we present a quantitative method to assess the skill of a set of numerical drifters. Using real-world drifting buoy trajectories, the chance can be calculated that the drifting buoys and the numerical drifters are drawn from the same distribution.

For assimilative models, where it is the objective for the model to represent the ocean state as accurately as possible, Barron et al. (2007) have developed a technique to compare drifting buoy

[*] Corresponding author. Address: Department of Physics and Astronomy, Institute for Marine and Atmospheric Research Utrecht, Utrecht University, Princetonplein 5, Utrecht, The Netherlands. Tel.: +31 302532909.
E-mail address: E.vanSebille@uu.nl (E. van Sebille).
[1] Present address: Department of Meteorology, University of Reading, United Kingdom.

1463-5003/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ocemod.2009.05.005

trajectories with the trajectories of numerical drifters. The authors seed numerical drifters at the locations where drifting buoys are observed and then calculate the deviation of model and in-situ paths as a function of time. However, many models are non-assimilative and for these models one-to-one comparison of buoys and numerical drifters is futile as the forcing is different between the model and drifting buoy trajectories. And even if the forcing is similar, nonlinearity leads to de-coupling (or rather de-timing) between the circulation and the forcing, and therefore, an increased error between observed and modeled trajectories. Verification should be done in a statistical sense, where the distribution functions of the two kinds of drifters are compared rigorously.

Lagrangian data is often used in examinations of relative and absolute dispersion. Such estimates of dispersion would be useful in quantifying important aspects of Agulhas circulation. For example, Drijfhout et al. (2003) identify dispersion through Rossby-wave radiation as a key factor in the decay of Agulhas rings. Lacorata et al. (2001) used Lyapunov exponents to characterize the drifter paths and assess the dispersion of drifting buoys. Manning and Churchill (2006) track the spread within drifter clusters in an alternate approach to estimating dispersion.

Drifter observations used within the present Agulhas study, however, are not well distributed for these type of methods, which analyze group characteristics of among multiple pairs or clusters of simultaneosly trajectories with initially small separation. Numerical simulations of drifter trajectories can be designed to support dispersion studies, but the validity of such studies requires that the simulated trajectories are representative of the true local circulation. The focus of the present study is to present a technique to assess whether the advection patterns in the model drifters agree with patterns in the real ocean. Model results that are shown to be sufficiently representative of observed characteristics could then be more credible in a subsequent study focused on dispersion characteristics.

Although drifting buoys have been deployed for over a decade now, and large numbers of buoys have been released, the total number of drifting buoys in a mesoscale region such as the Agulhas region is in the order of $10 - 10^2$. Numerical floats are seeded in quantities of $10^5 - 10^7$, many orders of magnitude larger. This small number of drifting buoy trajectories limits the ability to use standard statistical tools. A common χ^2-test, for example, requires histograms with at least five members in each bin. This confines the number of bins and consequently reduces the accuracy and strength of the method. A statistical test which is better suited for this problem is the two-sample Kolmogorov–Smirnov test, which does not require binning the data.

The method is applied to a set of experiments in the Agulhas system (De Ruijter et al., 1999; Lutjeharms, 2006), where numerical floats are continuously seeded in the upstream Agulhas Current and then tracked as they move through the Agulhas region. The highly nonlinear behavior of the flow in this region, with its dynamic retroflection and mesoscale eddies, serves as an ideal test case to investigate the strengths and weaknesses of the assessment method presented here.

2. The two-sample Kolmogorov–Smirnov test

To measure the agreement between the distribution functions of the numerical drifter data set and the drifting buoy data set, the two-sample Kolmogorov–Smirnov test (2KS-test) is used (Massey, 1951). The 2KS-test is designed to test the hypothesis that two data sets B (drifting buoys) and L (Lagrangian numerical drifters) are taken from the same underlying distribution. This underlying distribution does not need to be known. The two data sets have to be one-dimensional vectors of independent and identically distributed real numbers and they may have different lengths N_B and N_L, as the 2KS-test is also powerful when $N_B \ll N_L$. The 2KS-test starts out with formulating the null-hypothesis that B and L share an underlying distribution. After that, there are four steps (Fig. 1).

First, cumulative distribution functions $F_B(x)$ and $F_L(x)$ are constructed from the data sets B and L. These functions give the fraction of data below some value of the position x. They are zero below the minimum value in the data set and one above the maximum value. At each member of the (sorted) data set they increase with $1/N$. By construction, $F(x) = 0.5$ denotes the median of the data set.

Second, a test statistic is calculated. For the 2KS-test, this test statistic is the largest distance between $F_B(x)$ and $F_L(x)$:

$$D_n = \sup_x |F_B(x) - F_L(x)| \tag{1}$$

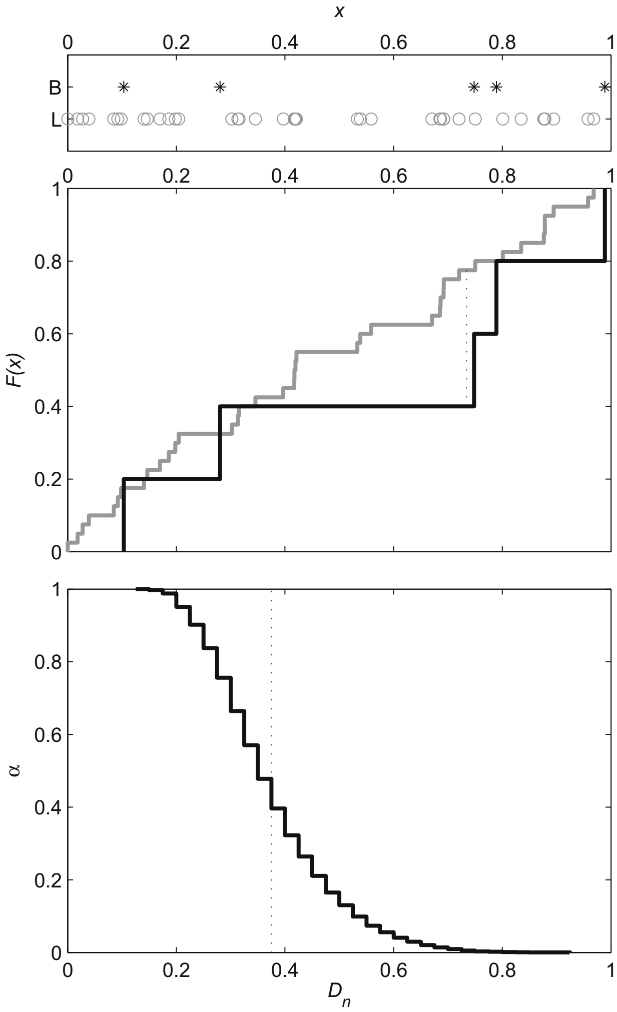

Fig. 1. An illustration of the two-sample Kolmogorov–Smirnov test. The test is performed using two random one-dimensional data sets B (asterisks) and L (circles), with $N_B = 5$ and $N_L = 40$ (upper panel), drawn from a uniform distribution. Cumulative distribution functions, the fraction of data points below some value x, have been computed from these two data sets (middle panel; black line for data set B and gray line for data set L). The test statistic D_n of Eq. (1) is denoted by the dotted line (with a value of 0.38). This test statistic is related to a confidence level α by a Monte Carlo process where D_n is calculated for 10^5 uniformly distributed data sets of similar N_B and N_L (lower panel). In this particular case the confidence level is 0.47, the value for α on the ordinate where the $D_n = 0.38$ line and the cumulative distribution function of all D_ns intersect. Since $\alpha > 0.05$, this leads to the (correct) conclusion that B and L are from the same distribution.

Other tests use different test statistics, such as the Cramér-Von-Mises test where the test statistic is the area between $F_B(x)$ and $F_L(x)$. However, these tests are not necessarily more powerful (Conover, 1980).

As a third step, a confidence level α is assigned to the test statistic D_n, given the data set lengths N_B and N_L. These two data set lengths are converted to one pseudo-length:

$$N = \frac{N_B N_L}{N_B + N_L} \tag{2}$$

after which the two-sample Kolmogorov–Smirnov test is similar to the ordinary Kolmogorov–Smirnov test. Note that for $N_B \ll N_L$, the length of the numerical drifter data set is unimportant as $N \approx N_B$.

The theory behind the 2KS-test states that, although the distributions of B and L may be unknown, D_n follows the Kolmogorov distribution. The transformation from D_n to α can be done using a look-up table (Sveshnikov, 1968; Conover, 1980), but here it is computed using a Monte Carlo simulation. In such a Monte Carlo simulation, a cumulative distribution function $F_{D_n}(\alpha)$ is acquired by repeatedly taking random samples of lengths N and calculating the test statistic. The advantage of using a Monte Carlo simulation over a look-up table is that it is much more accurate, at the cost of computing time.

Finally, the null hypothesis is rejected when α is below some value. In this paper, we use the 95% confidence interval. This leads to the decision rule:

$$\text{The model:} \begin{cases} \text{has no skill} & \text{if } \alpha \leqslant 0.05 \\ \text{might have skill} & \text{if } \alpha > 0.05, \end{cases} \tag{3}$$

which means that when $\alpha \leqslant 0.05$ it is more than 95% certain that the drifting buoy trajectories and numerical drifter trajectories do not share an underlying distribution and hence the model is not good. On the other hand, if $\alpha > 0.05$ it means that it is not certain whether the distributions of B and L are different. Although this technically only means that we can not say that the model is faulty, it will be used here as evidence that the model might have skill.

Note that in this formulation the 2KS-test only returns 'has no skill' or 'might have skill'. However, there is also information in the test statistic D_n and the confidence level α. They can be used for inter-model comparison. In addition, the cumulative distribution functions $F_B(x)$ and $F_L(x)$ can aid in subjective analysis as they reveal where model and reality diverge most.

The one-dimensional two-sample Kolmogorov–Smirnov test has been extended to two-dimensional data sets by Peacock (1983). The procedure is very similar in two dimensions, except for the conversion from B and L to $F_B(x,y)$ and $F_L(x,y)$. In two dimensions, there are four ways to define a cumulative distribution function, depending on where $F(x,y)$ is defined to be zero (Fig. 2). This is related to the possible orderings in x and y. As suggested by Peacock (1983), preliminary D_ns are computed for each of these four orderings, and the largest of these is selected as the representative D_n for the set, since that gives the smallest value for α,

$$D_n = \max\left(\sup_{x,y} |F_B(x,y) - F_L(x,y)|\right) \tag{4}$$

For a given N and D_n, the confidence level α is higher in the two-dimensional than in the one-dimensional 2KS-test (Fig. 3). This is probably because the average distance between N points in two dimensions is larger than in one dimension.

For sufficiently large data sets ($N > 10$), the 2KS-test is much more sensitive to changes in D_n than to changes in N (Fig. 3). If it is assumed that the data set is an unbiased subsample of the underlying distribution, so that D_n does not change when one data set member is added, the sensitivity of α is dominated by the change in N. For $N > 10$, the maximum sensitivity $|d\alpha/dN| = 0.05$.

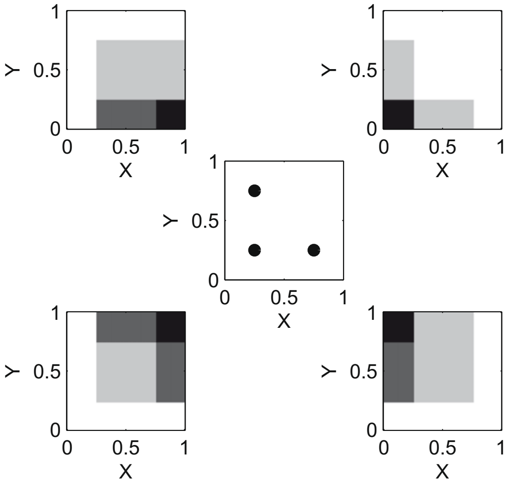

Fig. 2. The cumulative distribution functions (CDFs) that can be defined from a two-dimensional data set. Due to the orderings that can be made in the x and y dimensions, there are at least four different CDFs in two dimensions, where there is only one in one dimension. The middle panel shows an example data set where $N = 3$. The four corner panels show the four very different CDFs that result when the ordering is started in the respective corner of the (x, y)-domain. The color scale is such that white is zero and black is one. In the case of the two-dimensional 2KS-test, the ordering is chosen which results in the largest value of D_n.

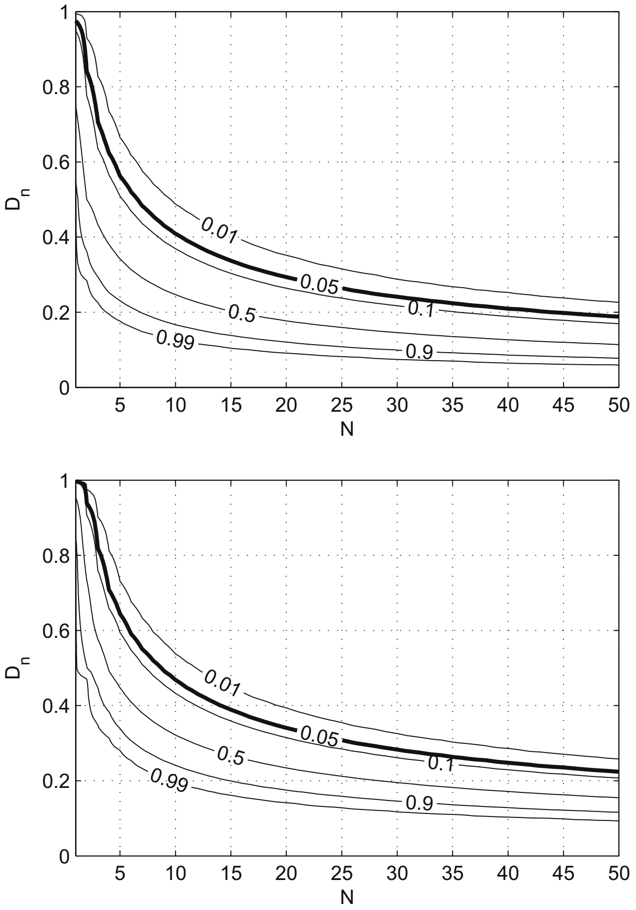

Fig. 3. Contour plots of the confidence level α as a function of N and D_n for the one-dimensional (upper panel) and two-dimensional (lower panel) Kolmogorov–Smirnov test. For a given N and D_n, the latter gives a higher confidence level.

Furthermore, the sensitivity around the critical $\alpha = 0.05$ value, which is the basis for the decision rule, Eq. (3), is always less than

0.01. That means that a decision will not have to be changed when $0.05 < \alpha < 0.06$ if one data point is added, under the assumption that D_n is constant. But even if the data set is extremely biased, the addition of one extra member to data set B can never change D_n by more than $1/N_B$, the height of each step in $F_B(x)$.

The advantage of the 2KS-test is that it is independent of a norm to compute the distance between the two data sets. Such a norm is required in the minimum spanning tree rank histogram method (Gombos et al., 2007), and this choice introduces subjectiveness into the method. The 2KS-test is, apart from a critical confidence level where the hypothesis is rejected, completely choice-free and thereby objective. Together with the ability of the 2KS-test to work for a large range of data set lengths, this makes the 2KS-test very appropriate for this oceanographic application.

3. The Agulhas region data

The 2KS-test is applied in the Agulhas region using drifting buoy trajectories as data set B and numerical drifter trajectories as data set L. The numerical drifter trajectories are obtained by seeding drifters in three different models, which means that there are actually three different data sets L.

The complexity and nonlinearity of the Agulhas system, where the Indian Ocean and Atlantic Ocean meet, makes it an ideal test case for the 2KS-test. The region is fed by three distinct sources: the Agulhas Current in the northeast, the South Atlantic subtropical gyre in the west, and the Antarctic Circumpolar Current in the southwest. The system is populated with mesoscale cyclones and anti-cyclones, which vigorously mix the water from these three source regions (Boebel et al., 2003). The float experiments are designed to determine the amount of Agulhas leakage, which is the water flowing from the Indian to the Atlantic Ocean in the warm upper-branch return flow of the thermohaline circulation (Gordon, 1985).

The three numerical data sets are from Lagrangian float experiments inside three different models: NCOM, ORCA, and AG01. The models vary in their ability to simulate the complicated Agulhas system dynamics, and this provides an opportunity to gauge the strength of the 2KS-test in this oceanographic context.

The 1/8° Global NCOM is an assimilative model in which satellite observations of sea surface height and temperature are used to derive synthetic profiles of temperature and salinity (Barron et al., 2006, version 2.5). Using seven years of model data in the Agulhas region (1998–2004), Lagrangian floats have been seeded daily according to volume flux in the Agulhas Current. Each float represents 0.1 Sv and the floats are tracked for two years. The total number of floats that is released at 30°S is 1.5×10^6.

The 1/2° global ocean sea-ice ORCA model (Biastoch et al., 2008c) is based on NEMO (Madec, 2006, version 2.3). It is forced with the Large and Yeager (2004) 6-hourly data set for wind and thermohaline forcing, over the period 1958–2004. The numerical floats are released at 32°S, employing the ARIANE package (Blanke and Raynaud, 1997). In the period 1992–2004, using the five day resolution model output, the floats are tracked for five years. In total, 1.3×10^6 floats are released.

The 1/10° AG01 model is a two-way nested grid inside the ORCA model, that spans the greater Agulhas region (20°W–70°E; 47°S–7°S) (Biastoch et al., 2008a,b). The two-way nesting procedure allows the AG01 model to both receive its boundary conditions from the base model and to update the base model (Debreu et al., 2008). The numerical float trajectories are computed in a similar way as in ORCA. In 37 years, 5.5×10^6 are released at 32°S.

The "truth", data set B, is a subset of the drifting buoy data set from the Global Drifting Buoy Data Assembly Center at the NOAA Atlantic and Oceanographic Meteorological Laboratory. The surface buoys have a drogue at 15 m depth. Richardson (2007) has used similar drifters to estimate Agulhas leakage and was able to identify some new features in the Agulhas region using all drifter trajectories in the domain. Here, the drifter data set has been limited to drifting buoys that flow downstream within the Agulhas Current. Since the numerical drifter release location is different between the models, two drifting buoy trajectory data sets are used. The trajectories start when the drifting buoys cross 30°S (NCOM) or 32°S (ORCA and AG01). Only that part of the trajectories is taken into account that is within the Agulhas region. These drifter trajectory boundaries are at 32°S and 40°E in the Indian Ocean, at 47°S in the Southern Ocean, and at 20°S and 20°W in the Atlantic Ocean. In total, the trajectories of 51 (NCOM) and 47 (ORCA and AG01) drifting buoys are used, in the period between 1995 and 2008.

In all three models, numerical floats are released throughout a large part of the water column. The drifting buoys, however, have a drogue at 15 m depth. Therefore, only the numerical floats in the upper 15 m of the models are used, and the models are only tested on their skill in the upper ocean (see also the discussion, Section 7). Technically, the numerical drifters released in AG01 and ORCA are not even drifters, as they are isopycnal and allowed to change their depth. If a float is within the upper 15 m, it is added to the drifter data set, irrespective of its depth history. However, we expect that the effect of resurfacing floats is minor.

4. Qualitative skill assessment

Although the goal of this article is to introduce a quantitative method for assessing the skill of an ocean model, we will start with qualitatively verifying the model results. This aids the interpretation of the results obtained later when the 2KS-test is applied.

The three models show very diverse behavior in the Agulhas region (Fig. 4). Of the drifting buoys in the real ocean approximately 25% end up in the Atlantic Ocean, and this is in agreement with recent estimates of Agulhas leakage (Doglioli et al., 2006; Richardson, 2007). In NCOM this fraction is much lower and this shows in the model trajectory density, which is very low in the Atlantic Ocean. However, the location and direction of the path taken by the Agulhas rings seems adequate. The Agulhas Return Current, at 37.5°E is better sampled.

The drifter density in ORCA reveals that the Agulhas leakage is directed too zonally, with the majority of the numerical drifters flowing westward after they round the Cape of Good Hope. This is an expression of the so-called Indian–Atlantic super-gyre (De Ruijter, 1982). In the Agulhas Return Current, at 37.5°E, a curious bi-partitioning can be seen. All drifting buoys flow eastward between 35°S and 42°S, but in the model there is an extra core around 33°S. A third discrepancy, which is to some degree also observed in NCOM, between the drifting buoys and the numerical drifters in ORCA is in the southward extent of the trajectories. The numerical drifters do not reach latitudes more southward than 41°S.

In AG01 the drifting buoy and numerical drifter distribution are much more in agreement, although the numerical drifters seem to enter the Atlantic on a too western course. No drifting buoys reach 0°E more southward than 25°S, but a vast amount of the numerical drifters from AG01 cross that longitude south of 30°S. Another discrepancy is in the return flow, where the distribution of numerical drifters is wider in latitude than that of the drifting buoys.

In summary, AG01 is qualitatively the best model. Although it has some deficiencies (too southward Agulhas leakage, too wide return flow), the area of maximum densities of the drifting buoys and numerical drifters coincide. The skill of NCOM is less, most notably in the fraction of drifters that get into the Atlantic Ocean (see also the discussion, Section 7). ORCA, at 1/2° resolution, seems to be the least skillful model with a preference for zonal flow.

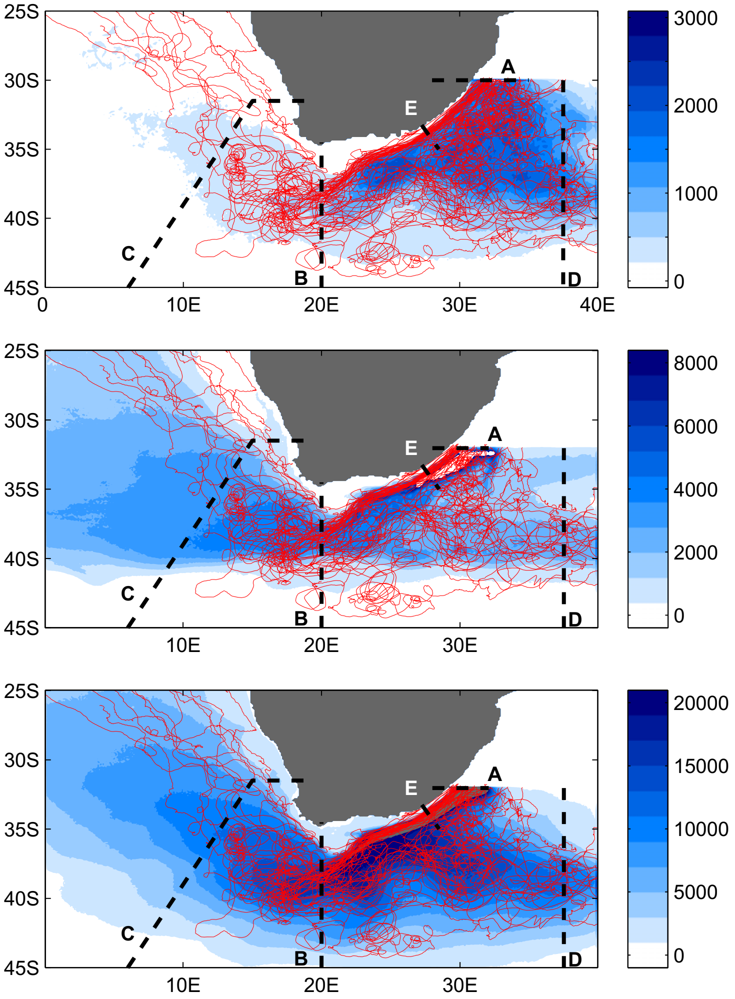

Fig. 4. The paths of the drifting buoys (red) after they have crossed the release latitude in the Agulhas Current and the density of numerical drifter trajectories (blue) for numerical drifters in the upper 15 m in NCOM (upper), the ORCA model (middle), and the AG01 model (lower). The black dashed lines denote the locations where the crossing positions of numerical drifters and drifting buoys are compared, the letters are for reference with Fig. 5 and Table 1.

5. Model validation along one-dimensional sections

The qualitative judgment of the skills of the three different models from the previous section can be quantified using the 2KS-test. For this, one-dimensional sections are taken at key locations in the Agulhas region. They are (A) the longitude of release at 30°S (NCOM) or 32°S (ORCA and AG01), (B) the highly variable retroflection at 20°E, (C) the Agulhas leakage at the GoodHope line (Swart et al., 2008), (D) the Agulhas Return Current location at 37.5°E, and (E) the Agulhas Current core attached to the continental slope as it passes Port Elizabeth and turns westward.

For each of the sections and all drifters, the position where the drifter crosses that section is added to the data set. Both the numerical and in-situ drifters may cross a section multiple times. If that is the case, the individual members of the data set are not independent anymore and the 2KS-test is formally not valid. To assure independence, each drifter can be in the data set only once. A way to resolve this is by adding the position of only the last crossing of each drifter to the data set. Using the first crossing instead of the last appears not to change the conclusions on model skill drawn below.

The data sets yield cumulative distribution functions similar to the one from Fig. 1. The qualitative skill assessment of Section 4 can be quantified using these cumulative distribution functions (Fig. 5), and the resulting confidence level α can be used to decide on the skill of the model over each of these five sections (Table 1).

At the latitude of release, both NCOM and AG01 perform well. In ORCA, on the other hand, the numerical drifter release locations are too far west, which implies that the modeled Agulhas Current is too confined to the African coast. This coastal confinement might be related to the absence of inshore cyclones that push the Agulhas Current offshore. These cyclones, called Natal pulses (Lutjeharms and Roberts, 1988), are not resolved on a 1/2° grid such as that in ORCA. Note that the drifter release locations in the ARIANE package, which is used in ORCA and AG01, are not continuous. Instead, drifters are only released in the center of the grid cells. Because the resulting distribution function is discrete, the transformation from test statistic D_n to confidence level α has to be performed using the finite sums method described by Conover (1980). This method is much more laborious than a Monte Carlo simulation but is exact for discrete distribution functions and yields confidence levels smaller than those for continuous distribution functions.

At 20°E, the section that cuts through the Agulhas Current retroflection, only AG01 might have skill. As also observed in the qualitative skill assessment, drifter trajectories do not get southward enough in ORCA. In NCOM, on the other hand, some numerical drifters are located too far southward.

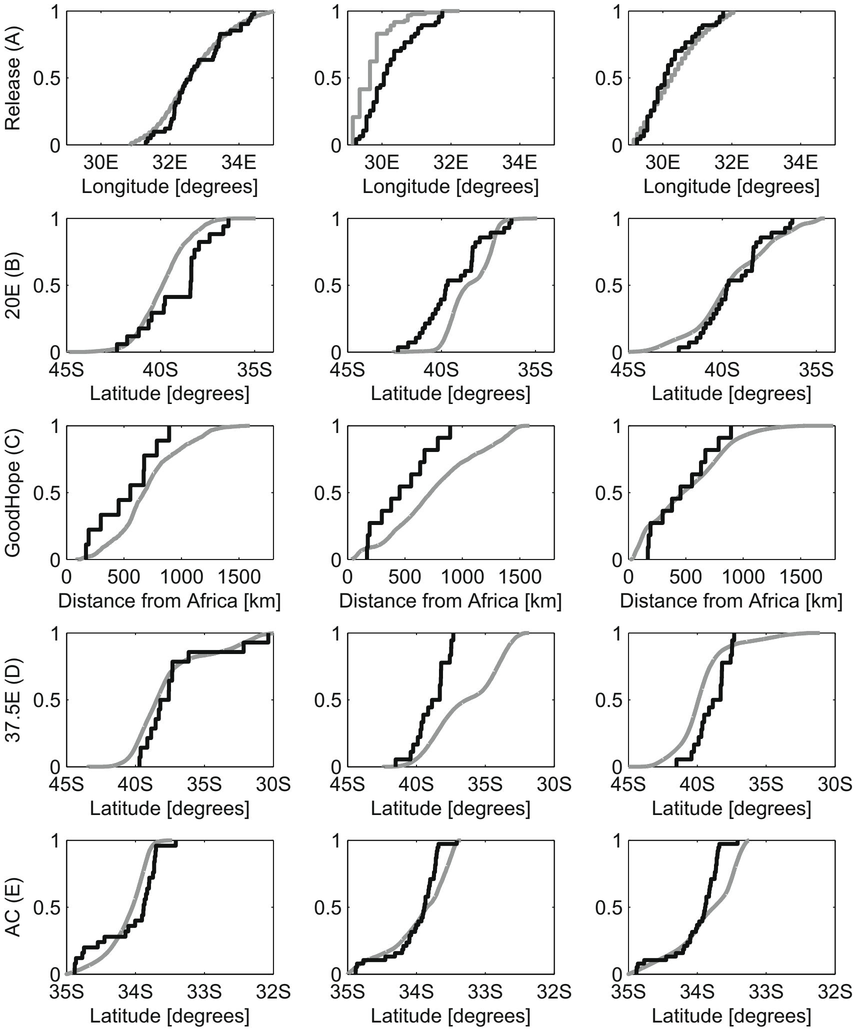

Fig. 5. The cumulative density functions $F(x)$ for drifting buoys B (black) and numerical drifters L (gray) for the five different sections depicted in Fig. 4 on each row. Results are shown from NCOM (left column), the ORCA model (middle column), and the AG01 model (right column). Since the numerical drifters in AG01 and ORCA are released at a different latitude than in NCOM, the drifting buoy data sets are also somewhat different. As the length of L is in the order of 10^5, the gray lines appear smooth. The confidence levels associated with these cumulative distributions functions are tabulated in Table 1.

Table 1
Confidence levels α for the two-sample Kolmogorov–Smirnov test applied to the five different sections shown in Figs. 4 and 5 for the three different models. According to the decision rule, Eq. (3), sections where $\alpha < 0.05$ are sections where the model does not have skill.

	Section	NCOM	ORCA	AG01
A	Release latitude	0.26	2.0×10^{-8}	0.21
B	20°E	1.7×10^{-3}	5.5×10^{-4}	0.79
C	GoodHope line	0.44	7.2×10^{-2}	0.47
D	37.5°E	0.49	6.2×10^{-6}	1.4×10^{-2}
E	Current core	6.4×10^{-3}	2.4×10^{-4}	7.4×10^{-7}

At the GoodHope line, both NCOM and AG01 might have skill, even though both models have drifter crossings too far offshore. But because there are only 11 drifting buoy crossings at this section, D_n is allowed to be larger before the decision has to be taken that the model has no skill (Fig. 3). Even ORCA might have skill, with a confidence level slightly higher than 0.05. One of the deficiencies of the one-dimensional 2KS-test is demonstrated here. NCOM severely underestimates Agulhas leakage, but because the crossing positions of the few drifters that do make it to the Atlantic Ocean are good, the model is designated skillful at the GoodHope line.

At 37.5°E, only NCOM might have skill. In ORCA the drifters cross too far northward, which is the expression of the bipartition also observed in Fig. 4. The annotation from this figure on AG01 at 37.5°E, that the spread of drifters is too wide compared to the drifting buoys, is confirmed in Fig. 5. Moreover, the median of the drifter crossings is more southward in AG01 than in the drifting buoy data.

According to the results tabulated in Table 1, none of the models have skill in the Agulhas Current as the numerical drifters are more coast-bound than the drifting buoys (Fig. 5). However, the difference is in the order of tens of kilometers which is mainly due to the details in the topography. All models use a land-mask which ends slightly too far northward and does not fully resolve the inner-shelf bathymetry. This causes a slight bias between the Agulhas Current core as sampled by the drifting buoys and in the models. This shows that one should be careful when applying the 2KS-test, especially if sections are very short.

Based on these five sections, it can be concluded that AG01 and NCOM are the best models, with possible skill at three sections. ORCA might have skill only at the GoodHope line. The highest confidence level α is found at 20°E in AG01, with only 20% chance that the numerical drifters and the drifting buoys are from a different distribution. This is above the critical confidence level of $\alpha = 0.05$ from the decision rule, Eq. (3).

The sensitivity of these results can be estimated by determining how the decisions would change if an extra drifting buoy crossing was added to the data sets of each section (Fig. 6). This figure denotes for each of the models and sections in Fig. 5 what its values for N and D_n are. Moreover, it shows the robustness of the skill decision since it divides the $N - D_n$ space in three regions depending on what can happen to the model skill decision if one new drifting buoy crossing is added to data set B. These are: a region where a model will never have skill when B is extended by one buoy; a region where a model might always have skill when B is extended by one buoy; and a region where the decision might have to be changed by the extension of B. As discussed in Section 2, adding one extra member to data set B can alter $F_B(x)$ by only $1/N_B$, and consequently the change in D_n is also at most $1/N_B$. Only the decision at the GoodHope line in ORCA could change in this worst-case scenario; all other decisions are immune to one extra drifting buoy crossing. Note that the dashed lines in Fig. 6 denote the maximum influence region. In reality, an extra buoy crossing would probably fall within the already found distribution and the change in D_n would likely be much smaller than $1/N_B$.

6. Two-dimensional model validation

In the previous section, it was concluded that both NCOM and AG01 might have skill at three of the five sections. However, from Fig. 4, it is clear that the numerical drifter trajectories in AG01 are in better agreement with the drifting buoy trajectories than they are in NCOM. This qualitative statement can be quantified using the two-dimensional 2KS-test, which yields a domain-wide measure of near-surface model skill.

One can not simply use the 2KS-test as described in Section 2, since the individual points that make up a trajectory are certainly not independent. This is because the location of a drifter at a certain moment is to a large extent determined by its former location. However, independence is required for the 2KS-test to be valid. To circumvent this problem of high interdependence of the data set, two ways are presented to adjust the two-dimensional 2KS-test.

6.1. Time-dependent confidence levels

The trajectory data sets are not independent because they contain information on the location of a drifter over a course of time. But this interdependence can be removed by taking only one position per drifter into account. To do this, the release of each drifter is synchronized to $t = 0$. Then, if the drifter positions are available at resolution Δt, the two-dimensional 2KS-test can be performed at every moment $t = n\Delta t$. This results in time series of the test statistics $D_n(t)$ and confidence levels $\alpha(t)$.

As time increases, the number of drifters in the model domain decreases because the drifters exit as they cross the domain boundaries. The two-dimensional 2KS-test is only valid for $N \gtrsim 10$ (Peacock, 1983). Therefore, the time series for α is trimmed to the moment that the number of drifting buoys in the model domain reaches 10. This occurs after 6 months in NCOM and after 5 months in ORCA and AG01. Using only the $\alpha(t)$ for the time that $N > 10$ yields mean confidence levels of 5.3×10^{-3} (NCOM), 1.9×10^{-2} (ORCA) and 0.11 (AG01). Using a 5 months window for NCOM instead of a 6 month window changes the mean confidence level to 4.2×10^{-3}. The conclusion must therefore be that, using this method and the decision rule, Eq. (3), only AG01 might possess skill when all drifter trajectories are taken into account.

6.2. Estimating the degrees of freedom

A second way to estimate the confidence level for the interdependent complete trajectories is by just ignoring the interdependence. Using the complete data sets, cumulative distribution functions can be calculated. From these cumulative distribution functions, a test statistic can be calculated just as in Eq. (4). Although technically the interdependence of the points in the data set prevents the 2KS-test from being valid, the test statistic does possess information.

This procedure yields values for the test statistic D_n of 0.25 for NCOM, 0.31 for ORCA, and 0.18 for AG01. If one assumes complete dependence of all data points on one trajectory, the number of degrees of freedom is just the number of trajectories. This is 51 for NCOM, and 47 for ORCA and AG01. Taking this amount of drifting buoys for N in the conversion from D_n to α leads to confidence levels of 8.0×10^{-3} (NCOM), 4.3×10^{-4} (ORCA) and 0.24 (AG01). This is an upper limit for the confidence level, as the relation between N and α is inverse so that more degrees of freedom lead to lower values of α (Fig. 3). From this upper limit, it must be concluded that ORCA and NCOM possess no skill in accordance with the decision rule, Eq. (3). The AG01 model might possess skill for $N = 47$, but the confidence level drops below 0.05 when $N = 82$.

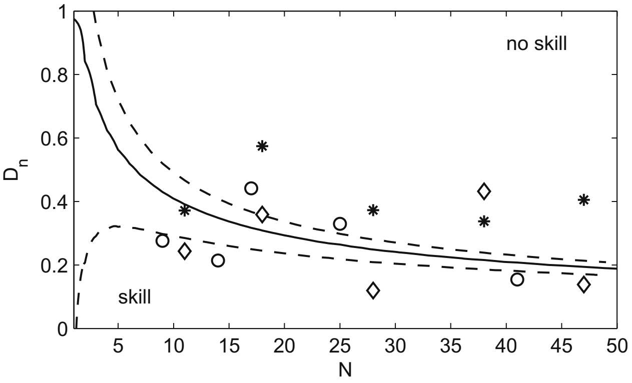

Fig. 6. Combinations of drifting buoy data set length N and test statistic D_n for the five sections of Fig. 4 and Table 1 for NCOM (circles), the ORCA model (asterisks) and the AG01 model (diamonds). The thick black line is the line where $\alpha = 0.05$, the divider between a model that lacks or might have skill according to the decision rule, Eq. (3). The area between the dashed lines denotes the region where in the worst-case scenario one extra drifting buoy could make α cross the 0.05 line. Only decisions inside this area are subject to change when an extra drifter is introduced.

7. Conclusions and discussion

We have applied the two-sample Kolmogorov–Smirnov test (2KS-test) to data sets of drifting buoys and numerical Lagrangian drifter trajectories. This test yields two numbers, the test statistic D_n and the confidence level α, which can be used to determine the skill of the model trajectories when the drifting buoy trajectories are taken as the truth. Moreover, the 2KS-test delivers a binary decision on the skill of the model. Depending on the value of α, the model either might have skill (when $\alpha > 0.05$) or has no skill (when $\alpha \leqslant 0.05$). These numbers come from the 95% confidence interval with which the hypothesis that numerical drifters and drifting buoys are drawn from the same distribution can be rejected.

The 2KS-test has been applied to three different models. The numerical drifter trajectories in the 1/2° ORCA model are so different from the drifting buoy trajectories that the model has skill in neither four out of five one-dimensional sections, nor in a two-dimensional sense. Only at the GoodHope line might the model have some skill, all be it not very robust. The 1/10° AG01 model might have skill in three of the five sections, and in the two-dimensional sense. The assimilative 1/8° NCOM has no skill in the two-dimensional sense, but it might posses skill in three of the five sections. This illustrates that, while a model may lack skill overall in a domain, it may have skill in certain locations. It also shows that the 11 buoy crossings through the GoodHope line might be too little for the 2KS-test to be useful, as this is the only section where the objective and subjective skill decisions differ.

Not only has AG01 the highest confidence levels in the Agulhas region, with an Agulhas leakage of 16.7 Sv it is also closest to estimates from many other studies (e.g. Doglioli et al., 2006; Richardson, 2007). NCOM, on the other hand, has a mean Agulhas leakage of only 1.5 Sv in addition to its lower confidence levels. This underestimation of Agulhas leakage is probably related to the mean location of the Agulhas Current retroflection, which is too far eastward in NCOM (Van Sebille et al., in preparation). However, it is unclear why a high resolution assimilative model is so underachieving in the Agulhas region. The low confidence levels of ORCA can probably be attributed to consequences of its course resolution. Due to the bad representation of the oceanic mesoscale in combination with the high explicit and numerical eddy viscosity the model shows a rather linear behavior. This results in an overestimation of Agulhas leakage of 32 Sv in ORCA (Biastoch et al., 2008a) and a relatively prominent supergyre (Biastoch et al., 2008b). But, as is also the case with these three models, diagnosing why a model lacks or might possess skill is much more difficult than assessing its skill.

In the implementation of the 2KS-test described here, model skill is a binary quantity: it is either 1 or 0. However, for model comparison the confidence level is probably a much better quantity. Using the magnitude of α also gets rid of the only choice one has to make when applying the 2KS-test: the choice for a critical confidence level. The presence of only one tuning parameter is one of the strengths of the 2KS-test. We have chosen for the 95% confidence interval for the critical value of α. Note that a larger confidence interval leads to a lower critical confidence level, which is defined in this way so that 'better' models have higher confidence levels, which is more intuitive than the other way around. Using the $\alpha = 0.05$ critical confidence level differentiated between a subjectively good model (AG01) and two subjectively bad models (NCOM and ORCA). But for confidence intervals between 90% and 98%, none of the decisions made have to be changed, either in the one-dimensional sections or the two-dimensional basin-wide assessment, except for the problematic decision on ORCA model skill at the GoodHope line.

The 2KS-test is here applied using drifting buoy trajectories as data set *B*. The disadvantage is that this confines the assessment to the skill of the upper 15 m. This is not a limitation of the method, but of the data set. In principle, other Lagrangian data sets (e.g. Argo floats or acoustic floats such as in RAFOS experiments) can also be used. There are two requirements for float data sets to be useful. Their time resolution should be high (a problem with Argo floats, which surface typically once a month), although this is not a requirement for two-dimensional skill assessment, and the number of trajectories should be sufficient (which is often a problem with RAFOS experiments).

The 2KS-test might even be applied to data sets beyond those of Lagrangian drifters or floats. In a Eulerian framework, the distribution of for instance model temperature at some grid-point could be compared to the distribution of temperature as obtained by a mooring. If one is interested in model-mooring validation of the complete distribution of some field, the 2KS-test can give a quick and objective measure of skill. However, if one is interested in assessing the variability, than an analysis of spectrum would be more suitable, as all temporal information is disregarded in the construction of the cumulative distribution function $F(x)$.

Acknowledgement

EvS is sponsored by the SRON User Support Programme under Grant EO-079, with financial support from the Netherlands Organization for Scientific Research, NWO.

References

Barron, C.N., Kara, A.B., Martin, P.J., Rhodes, R.C., Smedstad, L.F., 2006. Formulation, implementation and examination of vertical coordinate choices in the Global Navy Coastal Ocean Model (NCOM). Ocean Model. 11, 347–375.

Barron, C.N., Smedstad, L.F., Dastugue, J.M., Smedstad, O.M., 2007. Evaluation of ocean models using observed and simulated drifter trajectories: impact of sea surface height on synthetic profiles for data assimilation. J. Geophys. Res. 112, C07019.

Biastoch, A., Lutjeharms, J.R.E., Böning, C.W., Scheinert, M., 2008a. Meso-scale perturbations control inter-ocean exchange south of Africa. Geophys. Res. Lett. 35, L20602.

Biastoch, A., Böning, C.W., Lutjeharms, J.R.E., 2008b. Agulhas leakage dynamics affects decadal variability in Atlantic overturning circulation. Nature 456, 489–492.

Biastoch, A., Böning, C.W., Getzlaff, J., Molines, J.-M., Madec, G., 2008c. Causes of interannual-decadal variability in the meridional overturning circulation of the midlatitude North Atlantic Ocean. J. Clim. 21, 6599–6615.

Blanke, B., Raynaud, S., 1997. Kinematics of the Pacific Equatorial Undercurrent: an Eulerian and Lagrangian approach from GCM results. J. Phys. Oceanogr. 27, 1038–1053.

Boebel, O., Lutjeharms, J.R.E., Schmid, C., Zenk, W., Rossby, T., Barron, C., 2003. The Cape Cauldron, a regime of turbulent inter-ocean exchange. Deep Sea Res. II 50, 57–86.

Conover, W.J., 1980. Practical Nonparametric Statistics. John Wiley & Sons. 493pp.

De Ruijter, W.P.M., 1982. Asymptotic analysis of the Agulhas and Brazil current systems. J. Phys. Oceanogr. 12, 361–373.

De Ruijter, W.P.M., Biastoch, A., Drijfhout, S.S., Lutjeharms, J.R.E., Matano, R.P., Pichevin, T., van Leeuwen, P.J., Weijer, W., 1999. Indian–Atlantic interocean exchange: dynamics, estimation and impact. J. Geophys. Res. 104, 20885–20910.

Debreu, L., Vouland, C., Blayo, E., 2008. AGRIF: Adaptive grid refinement in Fortran. Comput. Geosci. 34, 8–13.

Doglioli, A.M., Veneziani, M., Blanke, B., Speich, S., Griffa, A., 2006. A Lagrangian analysis of the Indian–Atlantic interocean exchange in a regional model. Geophys. Res. Lett. 33, L14611.

Drijfhout, S.S., Katsman, C.A., De Steur, L., Van der Vaart, P.C.F., Van Leeuwen, P.J., Veth, C., 2003. Modeling the initial, fast sea-surface height decay of Agulhas ring "Astrid". Deep Sea Res. II 50, 299–319.

Gombos, D., Hansen, J.A., Du, J., McQueen, J., 2007. Theory and applications of the minimum spanning tree rank histogram. Mon. Weather Rev. 135, 1490–1505.

Gordon, A.L., 1985. Indian–Atlantic transfer of thermocline water at the Agulhas retroflection. Science 227, 1030–1033.

Griffies, S.M., Biastoch, A., Böning, C., Bryan, F., Danabasoglu, G., Chassignet, E.P., England, M.H., Gerdes, R., Haak, H., Hallberg, R.W., Hazeleger, W., Jungclaus, J., Large, W.G., Madec, G., Pirani, A., Samuels, B.L., Scheinert, M., Gupta, A.S., Severijns, C.A., Simmons, H.L., Treguier, A.-M., Winton, M., Yeager, S., Yin, J., 2009. Coordinated Ocean-ice Reference Experiments (COREs). Ocean Model. 26, 1–46.

Hetland, R.D., 2006. Event-driven model skill assessment. Ocean Model. 11, 214–223.

Lacorata, G., Aurell, E., Vulpani, A., 2001. Drifter dispersion in the Adriatic sea: Lagrangian data and chaotic model. Ann. Geophys. 19, 121–129.

Large, W.G., Yeager, S.G. 2004. Diurnal to decadal global forcing for ocean and sea-ice models: the data sets and flux climatologies, NCAR Technical Note, NCAR/TN-460+STR.

Lutjeharms, J.R.E., Roberts, H.R., 1988. The Natal Pulse: an extreme transient on the Agulhas current. J. Geophys. Res. 93, 631–645.

Lutjeharms, J.R.E., 2006. The Agulhas Current. Springer. 330pp.

Madec, G. 2006. NEMO ocean engine. Note du Pôle de modélisation, Institut Pierre-Simon Laplace (IPSL).

Manning, J.P., Churchill, J.H., 2006. Estimates of dispersion from clustered-drifter deployments on the southern flank of Georges Bank. Deep Sea Res. II 53, 2501–2519.

Massey Jr., F.J., 1951. The Kolmogorov–Smirnov test for goodness of fit. J. Am. Stat. Assoc. 46 (253), 68–78.

Peacock, J.A., 1983. Two-dimensional goodness-of-fit testing in astronomy. Mon. Not. R. Astr. Soc. 202, 615–627.

Richardson, P.L., 2007. Agulhas leakage into the Atlantic estimated with subsurface floats and surface drifters. Deep Sea Res. I 54, 1361–1389.

Sveshnikov, A.A., 1968. Problems in probability theory, mathematical statistics and theory of random functions. W.B. Saunders Company. 481pp.

Swart, S., Speich, S., Ansorge, I.J., Goni, G.J., Gladyshev, S., Lutjeharms, J.R.E., 2008. Transport and variability of the Antarctic Circumpolar Current south of Africa. J. Geophys. Res. 113, C09014.

van Sebille, E., Barron, C.N., Biastoch, A., Vossepoel, F.C., van Leeuwen, P.J., de Ruijter, W.P.M. 2009. Estimating Agulhas leakage from the Agulhas Current location. Ocean Sci, submitted for publication.