Anonymous Referee #2

General comments

In this paper the authors propose a classification method for determining ice types in Sentinel-1 SAR images. The structure and methodology are similar to those found in other studies, and overall the paper reads reasonably well. The study makes good use of recent published work by the authors on denoising Sentinel-1 SAR images, and improvements to the calculation of texture information in these images.

First of all, we would like to thank the reviewer for the positive evaluation and providing important comments. We hope that all your concerns will be cleared after reading our responses and modifications made to the manuscript. Please find below our answers (in green) and modifications (deleted in red and added in blue) to your comments/suggestions/questions.

Specific comments

To the best of my knowledge, this is the first study to examine the classification of Sentinel-1 SAR images for determination of sea ice types. These images are of great interest to the scientific and operational community. As the authors point out, the images are noisy, and the residual noise after the ESA correction is still significant. Certainly, the ability to classify ice types from such noisy images is of great use. However, I have difficulty following some of the claims made, in particular in the abstract and introduction. Primarily, I am not certain if it is clear to the authors that operational ice charts are generated manually, and contain significant bias and other possible errors of subjectivity. It is a little difficult to find information about this online, but the studies by Partington et al. (2003) and in the text by Johannessen et al. (2006) clearly state that the preparation of NIC charts (former reference) and AARI charts (latter reference) is through manual inspection of various sources of satellite imagery and other sources of data. Other studies (such as J. Karvonen, 2015) look at the accuracy of manual analyses by ice analysts. Training using a large volume of these charts would reduce operator-to-operator bias, but not the overall bias these charts are believed to contain since they are produced in the interest of marine safety. Based on this, the claim in the abstract and elsewhere that the use of ice charts allows training/testing data 'void of biased subjective decisions' should be revised.

Thank you very much for pointing an important issue. We revised the abstract and some parts in introduction as belows.

[Abstract]

A new Sentinel-1 image-based sea ice classification algorithm is proposed to support ~~automated~~ daily ice charting by using a machine learning-based model trained in a semi-automated manner. Previous studies mostly rely on manual work in selecting training and validation data ~~void of biased, subjective decisions~~. We show that the use of readily available ice charts from an operational ice services ~~allow to automate selelction~~ can reduce the amount of manual works in preparation of large amount of training/testing data. Furthermore, it reduces the inconsistent decisions in the classification algorithm by indirectly exploiting the best ability of the sea ice experts working at operational ice services.

…

[Section 1]

…

The use of public ice chart as training and validation reference data may help in solving the validation problem and enabling automation. The preparation of public ice chart is also through manual inspection of various sources of satellite imagery and other sources of data (Partington et al., 2003; Johannessen et al., 2006); however, training using a large volume of these charts would reduce operator-to-operator bias. The overall bias may exist since the public ice charts are produced in the interest of marine safety. Nevertheless, as the human interpretation available in the ice chart is currently considered as the best available information of sea ice (Karvonen et al., 2015), the best practice to make a sea ice type classifier is to train with the public ice chart so that the best knowledge of certified ice analysts is mimicked.

The 'novelty' of using ice charts in this way as training data should be clarified. These charts are fairly similar to the training data that was used for the sea ice type classification study by Zakhvatkina [2013], where homogeneous areas identified by trained ice analysts are used. Image analysis charts, which are very similar to daily ice charts with the exceptions that they are based only on the SAR imagery, are used directly as training data in the study by Wang et al. [2017]. In that study the ice concentration information was used directly in the same manner as ice type in the present study (the available charts were mapped to the SAR image latitude and longitude), however it was ice concentration information that was used, not ice type. These similarities should be discussed.

They are now included in the introduction as below.

[Section 1]

…

In ~~most~~many of the previous works on ice-water and/or sea ice classification (Soh and Tsatsoulis, 1999; Zakhvatkina et al., 2013; Leigh et al., 2014; Liu et al., 2015; Ressel et al., 2015; Zakhvatkina et al., 2017; Aldenhoff et al., 2018), the training and validation were done using manually produced ice maps. Although the authors claimed that the manual ice maps were drawn by ice experts, the selection of SAR scenes and interpretation can be ~~subjective~~inconsistent, and the number of samples were not enough to generalize the results because of the laborious manual work. Therefore, increasing objectivity is crucial, and automating the classification process is encouraged. The idea of training using SAR images and accompanying image analysis charts, which is raw interpretation of SAR images by trained ice analysts working at operational ice services, were tested for sea ice concentration estimation by Wang et al. (2017); however, such image analysis charts are not accessible to public.

Random forest classifiers are very popular at this time, and have been shown to be useful in many studies. To better motivate the present study, I suggest the authors compare their method to a multi-class random forest. In particular, the reference given for choosing the one-vs-all classification scheme as compared to multiclass problem is not closely related to the problem at hand. Did the authors try the multiclass method? Given that the motivation here is for operational implementation, it would be of interest to know if the mutliclass method performs similarly, and the computation time difference between the binary one-vs-all method and multiclass method.

Yes, we tried the multi-class random forest as well. The main reason for using one-vs-all scheme is to check the difference in feature importance per each of the classes. As the multi-class random forest gives a single feature importance, it is impossible to see the differences among the classes. The performance of the one-vs-all binary approach were slightly better than that of the multi-class method as shown in the table below, and this is in line with the results in Adnan and Islam, 2015. However, the computation times of the multi-class method were 1/3 and 1/2 compared to those of the binary one-vs-all method for the cases of 5- and 3-class.

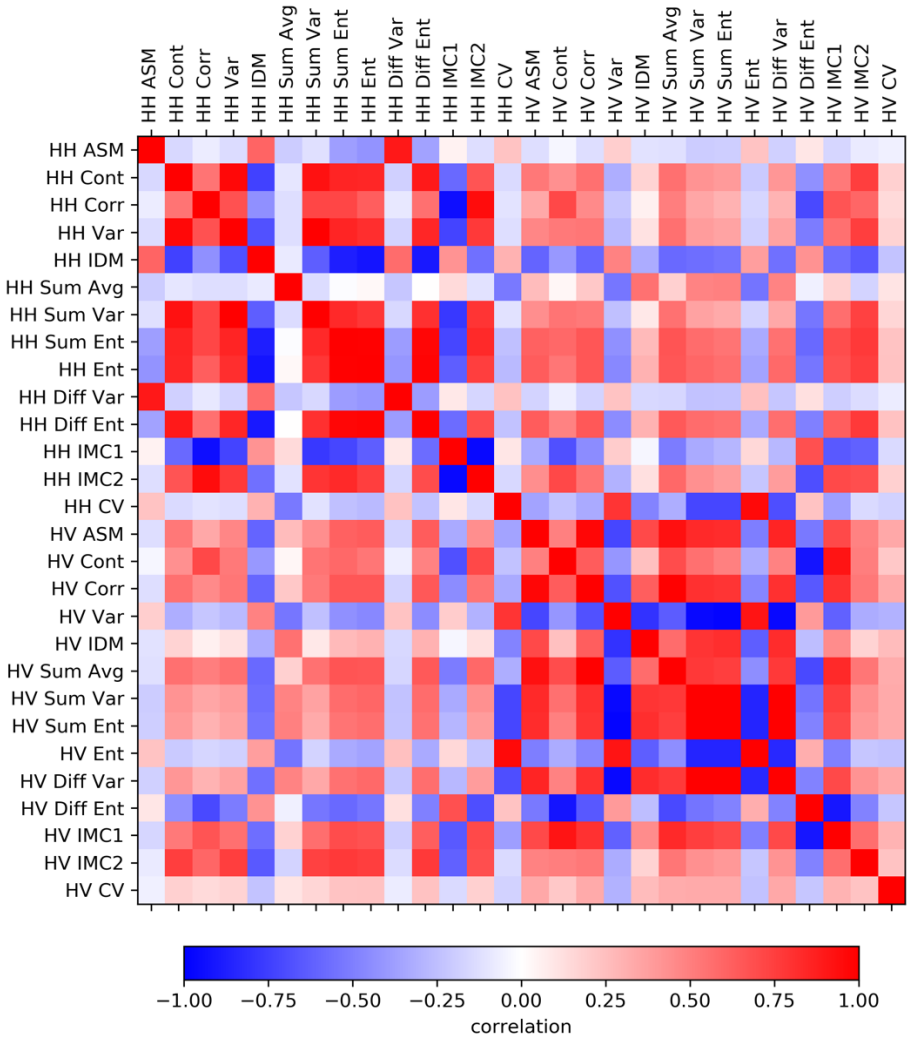
Feature configurations	FC1		FC2		FC3	
Number of classes	5 classes	3 classes	5 classes	3 classes	5 classes	3 classes
Overall accuracy (multi-class)	58.5	86.2	58.0	86.1	57.4	75.5
Overall accuracy (One-vs-all)	58.8	86.2	58.4	86.7	54.6	75.8
Cohen's kappa (multi-class)	0.66	0.79	0.66	0.79	0.52	0.53
Cohen's kappa (One-vs-all)	0.67	0.80	0.67	0.80	0.49	0.53

Adnan, M. N., and Islam, M. Z., One-Vs-All Binarization Technique in the Context of Random Forest, Proc. European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, Bruges (Belgium), 22-24 April 2015.

I have a similar question regarding the use of all Haralick texture features. How long did it take to calculate these features over the 64 grey levels used here? Are all features needed, or is it not relevant (in the sense that the additional time required and change in accuracy is not significant).

There are some high correlations between the features as shown in the heatmap below. For example, ASM and Diff Var, Cont and Var, Sum Ent and Ent have highly correlated each other. Removing some of them may not

lead to significant decrease in prediction accuracy, but the computational efficiency is out of the scope of this study. The computation time for extracting Haralick texture features per image is approximately five minutes in the given conditions (64 grey levels, 25 x 25 pixels of subwindow size) with an Intel i7 quad-core processor.



If the main contribution is to be the classifier itself, then a more careful examination of the method should be carried out. It would also be very interesting to see how the denoising methods they have developed lead to improved ice type classification. I am not sure if that would be difficult. Without this information, others are likely to attempt ice type classification without following rigorous denoising procedures. With this information, this piece of work could be a much stronger contribution to the sea ice community.

We conducted an additional test by following your suggestion, and the results are added to the revised manuscript. As shown in the table below, the textural denoising led to improved accuracies for all the classes except New ice.

[Section 3]

...

To see how the denoising step in Section 2.2.2 led to improvements in the classification accuracies, the same training and evaluation was conducted for the same dataset without applying the textural noise correction, and Table 4 shows the results. In both FC1 and FC2, the improvements in accuracies for young ice (+8.2-9.8%) and first-year ice (+9.2-11.6%) were most pronounced compared to those for open water (+1.7%) and old ice (+1.2-1.7%). On the contrary, a small decrease was observed for new ice (-2.8-4.7%). Nevertheless, the improvement in kappa (+0.05) demonstrates clear improvement in the overall classification result.

Table 4: Changes in classification accuracies before and after applying textural denoising

class	case								
	FC1			FC2			FC3		
	Thermal denoising only	Textural denoising applied	difference	Thermal denoising only	Textural denoising applied	difference	Thermal denoising only	Textural denoising applied	difference
OW	88.4	90.1	+1.7	88.9	90.6	+1.7	88.0	85.4	-2.6
NI	30.2	28.0	-2.8	27.7	23.0	-4.7	31.8	23.9	-7.9
YI	34.9	44.7	+9.8	36.2	44.6	+8.2	43.4	51.5	+8.1
FYI	29.3	38.9	+9.6	30.4	42.0	+11.6	38.0	47.0	+9.0
OI	91.5	92.7	+1.2	90.3	91.7	+1.4	75.2	66.3	-8.9
kappa	0.62	0.67	+0.05	0.62	0.67	+0.05	0.54	0.49	-0.05

In the end, it is found that the classification accuracies are higher when considering only three classes, first-year ice, muliti-year ice and open water. Could the authors add a little discussion to the conclusions as to if they envision a three-class or five-class operational implementation? If it is three-class, would they recommend using ice types from another sensor as training data? Some discussion on how the method is expected to work for other times of year should also be included.

We added relevant discussions to Section 3 and 4.

[Section 4]

...

Based on the results, we envisage that 3-class ice type classification from SAR imagery would be useful for making a global sea ice type product like EUMETSAT OSI-403-C (Aaboe et al., 2014) with higher spatial resolution.

[Section 3]

...

The proposed algorithm has several limitations as follows. First, the variations in radar backscattering and its corresponding image textures due to seasonal changes were not properly captured. Although the day of year was tested as a seasonality variable in the FC3 feature configuration, the result did not show any improvement. This is because day of year might not correspond to the same temperature, fluxes, and weather regimes.

page 6 - line 10 - Can the authors explain what they mean here by a 'sparse dataset' and why a dataset used for ice/water and ice types from SAR imagery would be considered a 'sparse dataset'? I am not sure I follow this line of reasoning.

The sentence was reworded as belows.

In the literatures about sea ice classification, the SVM was used often because by nature it works relatively well ~~for sparse dataset~~ when the number of datasets are small.

page 6 - line 32 - Why is the 'Richard's curve' chosen over a typical curve fit? Do the authors obtain more robust or interpretable results using this method? Please provide more context.

We added more explanations.

> ...
> Classification scores with values ranging from 0 (worst performance) to 1 (best performance) are evaluated for each node of the grid and are interpolated between the nodes by curve fitting. ~~t~~The Richard's Curve (Richard, 1959) was used as the fit model because it allows easy estimation of the model's maximum value.

page 7 - If I understand correctly, the authors manually selected 57 image (or do the authors mean scenes here?) for training and testing from a set consisting of 958 images (or again is this scenes)? Can they say something about these 57? Are they from similar geographic regions? times of year? specific features? Going through 958 images manually to choose a training data set is not automated. Using an ice type product generated in an automated manner from another sensor (for example open water/FYI/MYI from passive microwave data or scatterometer data) could provide an automated workflow.

If you mean the image-subimage things, it is SCENES here. As in Section 2.1, a total of 958 scenes were acquired, and the selected 57 scene are from various geographic region within the study area and various time of year. Using ice type product from passive microwave data or scatterometer data cannot help the image selection procedure due to the large difference in spatial resolution of them and SAR. Regarding the automation issue, we clarify throughout the revised manuscript that the developed algorithm is "semi-" automated.

> **[Sections 2.2.6]**
> **…**
> ~~To automate image selection for training, a good ice/water classifier for SAR image is needed.~~ In order to automate image selection, the ice edges in SAR images needs to be identified first. Since ~~even such a simple binary~~ an ice/water classifier has not been well developed yet for Sentinel-1, the image selection procedure has to be done manually in the beginning. However once a classifier is generated with high accuracy, it can be used to automate the procedure, then the whole process in the proposed scheme will be fully automated. This is why the proposed algorithm is named "semi-" automated for now. Nevertheless, the manual selection is done by visual inspection of ice-water boundaries overlaid on SAR images. The ice-water boundary can be extracted easily from the reprojected ice chart by selecting the pixel borders of open water class. Then the SAR backscattering image contrasts across the ice-water boundaries are examined both in HH- and HV-polarization because the image contrast between ice-water is larger in HV but smooth level ice is better recognizable in HH.

page 8 and Figure 6 - What method was used to determine the feature importance score and why was this method chosen? How is this score calculated?

> **[Sections 2.3]**
> **…**
> For each sub-classifier, each of the texture features has different weight in decision making. The fraction of the samples that each of texture features contribute to can be used to compute the relative importance of the features, and the averaged estimates of them over several randomized trees serve as an indicator of feature importance (Louppe, 2014). The feature importance for the sub-classifiers is presented in Figure 6.
>
> Louppe, G.: Understanding random forests: From theory to practice, PhD Thesis, U. of Liege, 2014.

Technical comments

1) abstract - overall accuracies vs. overall accuracy - Be consistent in your use of plurals here

Corrected.

2) abstract - In what way would this work support automated ice charting? Were the authors thinking that fewer operational (manual) charts would need to be produced? Clarification of this point would be helpful.

Revised.

> A new Sentinel-1 image-based sea ice classification algorithm is proposed to support daily ~~automated~~ ice charting by using a machine learning-based model trained in a semi-automated manner.

3) page 1 - line 12 - 'In most of the previous works...'... please provide a few references in this sentence to the works you have in mind.

References added.

> In ~~most~~ many of the previous works on ice-water and/or sea ice classification (Soh and Tsatsoulis, 1999; Zakhvatkina et al., 2013; Leigh et al., 2014; Liu et al., 2015; Ressel et al., 2015; Zakhvatkina et al., 2017; Aldenhoff et al., 2018), the training and validation were done using manually produced ice maps.

4) page 1 - lines 12-15 and lines 20. I reiterate my earlier point. Ice charts are generated manually by trained analysts. Although they are available in the public domain, and this means using these charts directly relieves the individual designing the classification algorithm from the 'laborious' and possibly biased process of manually choosing training and testing data, it does not enable an automated workflow.

Revised.

> A new Sentinel-1 image-based sea ice classification algorithm is proposed to support ~~automated~~ daily ice charting by using a machine learning-based model trained in a semi-automated manner. Previous studies mostly rely on manual work in selecting training and validation data ~~void of biased, subjective decisions~~. We show that the use of readily available ice charts from an operational ice services ~~allow to automate selelction~~ can reduce the amount of manual works in preparation of large amount of training/testing data. Furthermore, it reduces the inconsistent decisions in the classification algorithm by indirectly exploiting the best ability of the sea ice experts working at operational ice services.
> ...

5) page 1 - line 20 - Again, ice charts are generated by humans. They contain human error. They are often produced under a strong time constraint, and in the interest of marine safety (the latter point meaning they likely contain bias to ensure safety).

Revised.

> **[Section 1]**
> ...
> The use of public ice chart as training and validation reference data may help in solving the validation problem and enabling automation. The preparation of public ice chart is also through manual inspection of various sources of satellite imagery and other sources of data (Partington et al., 2003; Johannessen et al., 2006); however, training using a large volume of these charts would reduce operator-to-operator bias. The overall bias may exist since the public ice charts are produced in the interest of marine safety. Nevertheless, as the human interpretation available in the ice chart is currently considered as the best available information of sea ice (Karvonen et al., 2015), the best practice to make a sea ice type classifier is to train with the public ice chart so that the best knowledge of certified ice analysts is mimicked.

6) page 3 - line 25 - 'ice edge determined from AMSR-E' - an ice edge cannot be determined from AMSR-E without using an algorithm. Which algorithm was used? Please revise.

Revised.

> Heinrichs et al. (2006) reported that the ice edge determined from the AMSR-E~~, which is a~~ passive microwave radiometer data~~,~~ using the isoline of 15% concentration matches best the ice edge determined from RADARSAT-1~~, which is a C-band HH polarization~~ SAR data using visual inspection.

7) page 3 - lines 28-29 - I don't know what the authors mean by 'has a precision of decimals'.

Revised.

> Note that ice concentration label in the SIGRID-3 format ~~has precision of decimals~~ is assigned in increments of 10%.

8) page 3 - line 26 - Similar comment regarding the ice edge determined from SAR - a methodology must have been used to get this ice edge. Was it visual inspection, or another method? Please revise.

Revised.

> Heinrichs et al. (2006) reported that the ice edge determined from the AMSR-E~~, which is a~~ passive microwave radiometer data~~,~~ using the isoline of 15% concentration matches best the ice edge determined from RADARSAT-1~~, which is a C-band HH polarization~~ SAR data using visual inspection.

9) page 4 - line 3 - better than what?

Revised.

> Comparing the original SoD i~~I~~n the top left panel with the processed SoD in the bottom left panel, it is clear that the ice edge of the processed SoD match better with the SAR backscattering images.

10) page 5 - line 2 - wording is not specific - many of the previously developed methods - methods for what? These references are a mixture of ice/water and ice type classification studies. These two tasks are different from the perspective of a computer algorithm. Also, a reference to Shokr [1991] should be included.

Revised and added reference.

> Like many of the previously developed sea ice type classification methods (Shokr, 1991; Barber and LeDrew, 1991; Soh and Tsatsoulis, 1999; Deng and Clausi, 2005; Zakhvatkina et al., 2013; Leigh et al., 2014; Liu et al., 2015; ~~Karvonen, 2017; Zakhvatkina et al., 2013, 2017~~), the proposed approach starts from gray level co-occurrence matrices (GLCM) calculation.

11) page 5 - lines 9-11 - I don't understand this sentence, what is averaged for multiple distances, and what is the normalized GLCM?

Normalized GLCM is the GLCM divided by the sum of all elements, representing probability of co-occurrence. As there are multiple normalized GLCMs, one for each of the co-occurrence distances, the averaged values were used to reduce the dimensionality of the data to analyze.

12) page 5 - line 5 - direction? or should this be orientation?

Revised. It should be orientation in the context.

13) page 5 - line 12 - The term spatial resolution is not clear. Some authors consider this the scales that are resolved. It may be better to state that the spacing between the GLCM texture feature windows is 1km? (or please reword if I am not interpreting this point correctly).

Revised.

> In this study, we set *w* as 25 so that the ~~spatial resolution~~grid spacing of the result of texture analysis is 1 km.

14) page 5 - It would be nice to have the Haralick features listed in a table, and to provide a brief rationale for including all of them in the study. Information as to how long it took to calculate these features using 64 grey levels for their set of imagery is also important.

As the usefulness of GLCM-based texture features for sea ice classification has been demonstrated in literature (Shokr, 1991; Soh and Tsatsoulis, 1999; Deng and Clausi, 2005; Zakhvatkina et al., 2013; Leigh et al., 2014; Liu et al., 2015) and the Haralick features include most of them, it might be not necessary to list all the features in the manuscript. The computation time for extracting Haralick texture features per image is approximately five minutes in the given conditions (64 grey levels, 25 x 25 pixels of subwindow size) with an Intel i7 quad-core processor.

15) page 5 - the number of Haralick features is referred to inconsistently as 13 on line 7 and 26 on line 25. The 26 is likely just accounting for the two polarizations, but the two should be referred to in a consistent manner. Similarly on page 7 lines 22-23, please use either 'Haralick texture features' or 'texture features' consistently when describing the three classifiers.

Revised.

In addition to the ~~26~~13 Haralick features, the coefficient of variation (CV) which is reported as useful feature for ice-water discrimination (Keller et al., 2017) is included.
…
We trained three RF classifiers with different feature configurations: i) FC1: ~~texture features from~~ Haralick texture features and CV, ii) FC2: Haralick texture features, CV, and incidence angle, iii) FC3: Haralick texture features, CV, incidence angle, and day of year.

16) page 6 - line 16 'they are' - what are 'they'? Is this the number of operations?
Revised.

For the SVM, ~~they~~the number of operations are $O(n^2p + n^3)$ and $O(n_{sv}p)$ for training and prediction while for RF, $O(n^2pn_{tr})$ and $O(n_{tr}p)$, respectively, where n is the number of samples, p is the number of features, n_{sv} for the number of support vectors, and n_{tr} for the number of trees.

17) page 7 - If a binary ice/water classifier is 'simple' (line 6), why are the authors starting with ice type classification? I suggest this be reworded.
Revised.

Since even ~~such~~ an ~~simple binary~~ ice/water classifier has not been well developed yet for Sentinel-1, the image selection procedure has to be done manually in the beginning.

18) page 8 - lines 20-21 - The sentence starting with, 'Since the training and test datasets were extracted from the same…' I find a little out of place. With this placement, it seems like it is trying to account for the results from FC2 and FC3. It might be better to start this one with 'When the evaluation is carried out with the 2018 data, the training and test datasets….'
Revised.

19) page 9 - line 32 - 'capturing' should be 'to capture'
Corrected.

20) page 9 - line 32 -'more details' - as compared to what?
Corrected.

21) Figure 3 - Could the authors provide some information in the text as to what the map of partial concentration is (top right). Is this the partial concentration of the dominant ice type for the given polygon?
Revised.

22) Figures 3,7,8 and 9 should have geolocation data provided.
Added geolocation grid to Figure 7 and Figure 8. In Figure 3 and Figure 9, the geolocation information may be irrelevant for understanding the contents.

23) all numbers less than ten should be written out in words, eg., 3 -> three
Corrected

References

A comparison between high-resolution EO-based and ice analyst-assigned sea ice concentrations, Juha Karvonen and others, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 8(4):1-9, 2015.
Evaluation of second-order texture parameters for sea ice classification from radar images, Mohammed E. Shokr, Journal of Geophysical Research, 96(C6),10,625-10640, 1991.
Late twentieth century Northern Hemisphere sea-ice record from the U.S. National Ice Center ice charts, Kim Partington, Tom Flynn, Doug Lamb, Cheryl Bertoia and Kyle Dedrick, Journal of Geophysical Research, 108(C11), doi:10.1029/2002JC001623, 2003.
Remote sensing of sea ice in the Northern Sea route: Studies and applications, Ola M. Johannessen and others, Springer Science and Business Media, 472 pages, 2006.
References above are now in the reference list of the revised manuscript.