Chapter 6

Discussions and future work

6.1 Discussion

The empirical relationship between pCO_2 and other ocean properties is investigated in this dissertation. Furthermore, the sampling of a small percentage of the available data points to determine the empirical relationship is investigated. Both least square curve fitting and RBF interpolation is used to determine the empirical relationship. D-optimal sampling is used to judicially sample a subset of points from the complete data set. The results are shown in Chapter 5. The results obtained in this dissertation are discussed in this section.

When comparing the results obtained with least squares curve fitting of the linear equation (with and without latitude), it can be seen that there is not much improvement in the error when the latitude is included. The RMS error of the 200 D-optimally sampled points remains approximately 20.30 μatm with latitude excluded or included as variable. The same results is observed for the quadratic equation where the RMS error of the 200 D-optimally sampled points for the quadratic equation remains approximately 18.2 μatm for the latitude excluded or included. For the cubic and fourth order equations however, the error of the estimation of pCO_2 is improved with the addition of latitude. For the cubic curve fit, the RMS error is improved from 17.94 μatm with latitude excluded as variable to 13.52 μatm with latitude included. For the fourth order curve fit, the RMS error of 200 D-optimally sampled points is decreased from 15.39 μatm with latitude excluded as variable to 8.797 μatm with latitude included. It can thus be seen that in the least squares curve fitting the addition of latitude as variable decreases the error in the estimation of the pCO_2 significantly.

As the order of the equation that is fitted to the data increases, the error on the estimation of the

pCO_2 decreases. The RMS error is reduced from 20.27 μatm for 200 D-optimally sampled points with the linear curve fit and latitude excluded as variable, to 15.36 μatm for 200 D-optimal sampled points with the fourth order curve fit and latitude excluded as variable. Similarly, for latitude included as variable, the RMS error for 200 D-optimal sampled points is reduced from 20.30 μatm for the linear curve fit to 8.797 μatm for the fourth order curve fit. The restriction in terms of the number of terms that can be used in the equation should be used in order to determine which equation will yield the smallest errors for this application.

The radial basis function interpolation estimates the pCO_2 with a smaller error than the fourth order equation, which is the equation with the smallest error on the estimation with the least squares optimization in this dissertation. The RMS error for 200 D-optimally sampled points with the fourth order curve fit with latitude excluded as variable is 15.36 μatm compared to a RMS error of 9.617 μatm for the 200 D-optimally sampled points with the RBF interpolation with latitude excluded as variable, without optimal scaling. Similarly, with latitude included as variable, the RMS error of the fourth order curve fit for 200 D-optimal sampled points is 8.797 μatm, compared to the RMS error of 6.761 μatm for the RBF interpolation for 200 D-optimal sampled points without optimal scaling.

The inclusion of latitude also reduces the error in the RBF interpolation significantly. The RMS error of the RBF interpolation without latitude is 9.617 μatm for 200 D-optimal sampled points, compared to the RMS error of 6.761 μatm for 200 D-optimal sampled points with latitude included as variable. It can be seen from the RBF results that latitude improves the error of the estimation of the pCO_2. Since the radial basis function is an interpolation function, the inclusion of latitude allows for the interpolation to implicitly take into account the regions that exist in the ocean. The compact support RBFs allows for the interpolation to take into account the regions and the variability of the relationship between the variables in different regions.

By optimally scaling the variables in the RBF interpolation, an even smaller error on the estimation is obtained. The RMS error, for the case where latitude is excluded as variable, is reduced from 9.617 μatm without the optimal scaling to 9.012 μatm with the optimal scaling for 200 D-optimal sampled points. For the case where latitude is included as variable, the RMS error, for 200 D-optimal sampled points, is improved from 6.716 μatm without the optimal scaling to 4.065 μatm with the optimal scaling. From the scaling in the case where the latitude is included as a variable, it can be seen that the relationship between the variables differ regionally, as the latitude is given a bigger scaling than the other variables. The latitude is scaled between 0 and 24, whereas the Temperature is scaled between 0 and 3, the MLD is scaled between 0 and 0.5 and the Chl is scaled

between 0 and 1.8. As a result, the radius of one that is specified in the compact support functions will result in more latitudinal regions considered when the compact support function interpolation is used. This means that even if the same value for T, MLD and Chl is obtained at two different points in the ocean, the position will still play a role in the estimation of the pCO_2 since the empirical effect that each of these variables have on the pCO_2 differs in different regions. The optimal scaling is significant in the RBF interpolation in order to make sense of the radius of one that is specified in the compact support functions implemented.

From the results is can be deduced that latitude is an important variable to include in the investigation of the empirical relationship between pCO_2 and the ocean variables. By doing this, no knowledge of the positions of the regions in the ocean is needed in order to obtain an accurate model. The latitude that is included as variable in the empirical relationships already, implicitly, takes the different regions into account.

Furthermore, it can be seen that in the least squares optimization, for a small number of points selected from the data set to use as subset to do the optimization, the D-optimal sampling improves the error on the estimation. This is helpful when a much larger data set is used to establish the empirical relationships in the entire Southern Ocean. When taking into account the entire Southern Ocean, more than 8 million data points are available, and we only want to use a subset of this to determine the empirical relationship between the ocean variables and pCO_2. It is shown in this dissertation that even when using a small number of points, the least square optimization can be done on the data in order to obtain an empirical estimation of the pCO_2 when the points are sampled using D-optimal sampling. The D-optimal sampling is also a more robust sampling method as can be seen from the 95% confidence interval and the standard deviation on the coefficients obtained for 100 different runs of the curve fitting.

For RBF interpolation, the points that are used in the data set to construct the interpolation function are also of importance in reducing the estimation error of pCO_2. When the subset of points that are selected for constructing the interpolation function is an accurate representation of the behaviour of the variables in the ocean, then the error on the prediction from this interpolation function is improved. The D-optimal sampling in this case yields smaller errors on the estimation than the random sampling.

As the number of points that are sampled (for the subsets of points from which the empirical relationships are determined) are increased, the prediction error is improved. Computational resources have to be taken into account when considering the amount of points chosen for the least squares optimization and the RBF interpolation. Another factor that needs to be taken into account in

the RBF interpolation is that the number of points selected to determine the coefficients for the interpolation function is also the amount of terms in the interpolation equation. The restrictions on the amount of terms in the equation need to be considered when deciding on the number of points sampled.

For the RBF interpolation, there is an optimal combination of factors that improves the estimation error to a great extend. Firstly, the latitude included as variable improves the estimation. Secondly, the D-optimal sampling yields smaller estimation errors than the random sampling. Thirdly, the optimal scaling greatly reduces the estimation error by scaling each variable such that the variability in each variable is taken into account. The RMS error of 4.065 μatm for the 200 D-optimally sampled points with latitude included as variable and the optimal scaling of the variables is almost a 1% error on the estimation of the pCO_2. The maximum over-estimation of 39.32 μatm and maximum under-estimation of -36.34 μatm in this case is approximately a 10% error on the estimation of the pCO_2 which is less than the error of 15% that was required for this project.

The hypothesis that an empirical relationship can be obtained is verified as is shown by both the least squares optimization and the RBF interpolation. An empirical relationship in the form of an equation can be found for both methods. Furthermore, it is verified that, by using D-optimal sampling, a subset of the data points can be selected to determine the empirical relationship. The D-optimal sampling yields a subset of points that allows for an empirical relationship that accurately estimates pCO_2 in the ocean.

The equations obtained in this dissertation is only for the stretch of the ocean that is covered by the SANAE cruise. In order to obtain an equation that can be used on the entire ocean, the methods discussed in this dissertation can be applied to data of the entire Southern Ocean. It can be concluded from the results in this dissertation that it is possible to use classical techniques to find an empirical relationship between pCO_2 and other ocean variables. The results also show that the pCO_2, in this data set, can be estimated with reasonable accuracy using both the fourth order curve fit as well as RBF interpolation. The D-optimal sampling can be successfully used to reduce the prediction error by sampling the optimal subset of points from the data set.

6.2 Future work

This dissertation is part of initial research on the estimation of Southern Ocean pCO_2. Taking these results as initial an investigation, more work can be done in order to improve the estimation of the pCO_2 in the Southern Ocean. Some recommendations are formulated here.

- The work in this dissertation is only based on a single ocean cruises over a period of 10 days. Some future work could include using the techniques used in this dissertation to extend the region over which the pCO_2 estimations are made. Essentially the entire Southern Ocean will be used as the region over which the pCO_2 estimations will be made.
- When the data set is extended over a larger ocean area, the variability of the variables and the relationship between the variables in a longitudinal direction can also be taken into account.
- The seasonal variability of the oceans can be included in these models to predict pCO_2 since the relationship between the variables can also change seasonally.
- For the least squares optimization, equations of higher order may be applied in order to get improved results for the estimation of pCO_2 depending on the restrictions on the number of terms in the equation.
- Other regression analysis methods can also be used to fit the data.
- The genetic algorithm's performance can possibly be improved if other operators and mutation and crossover probabilities are used. These can possibly be adjusted to improve the results for the D-optimal sampling.
- The D-optimal sampling can be performed by using other optimization methods (than the genetic algorithm) that may yield more optimal results for the D-optimal sampling.
- The optimal scaling that is performed for the radial basis function interpolation needs to be done using a more robust optimization method. This optimization to decrease the error by improving the scaling of the variables could be a whole new study by itself.
- Other interpolation methods can also possibly yield improved results when applied to the Southern Ocean data sets.