Response to Referee #1 Comments on Gemayel, E., Hassoun, A.E.R., Benallal, M.A., Goyet, C., Rivaro, P., Abboud-Abi Saab, M., Krasakopoulou, E., Touratier, F., Ziveri, P., 2015: Climatological variations of total alkalinity and total inorganic carbon in the Mediterranean Sea surface waters. Earth Syst. Dynam. Discuss. 6 (2), 1499-1533. 10.5194/esdd-6-1499-2015

P.S: Original referee comments are in normal font; our replies are in italics. Intended changes to text are shown in bold font.

Gemayel et al. present an interesting study regarding the sea surface total alkalinity and total inorganic carbon in the Mediterranean Sea. To-date our knowledge regarding the carbonate system is limited due to the sparsity of available observations, hence I very much appreciate the effort of the authors to gather available observations and perform this basin scale study. The authors investigate the spatial distribution as well as seasonal variability and nicely explain their findings. The manuscript is well structured, well written and nicely relates the findings of this manuscript to previous studies. I do however believe that the authors need to substantially improve their currently too short methods section. Please find specific points below in the major and minor comments sections.

We would like to thank the referee for their thorough comments, suggestions and criticisms. The points raised by them helped us to improve our manuscript.

Major comments:

The authors need to add more detail regarding the 10-fold cross validation technique. E.g: how are the subsets (training, testing) chosen? randomly?

A more detailed description regarding the 10-fold cross validation technique was added. We modified this section according to:

This model validation technique is performed by randomly portioning the dataset into 10 equal subsamples. One subsample is used as the validation data, and the 9 remaining subsamples are used as training data. The cross validation process is then repeated 10 times, with each of the 10 subsamples used exactly once as the validation data. In this manner, all observations are used both for training and validation, and each observation is used for validation only once.

How is the data distribution between training and testing established? On page 1504 last line the authors report 375 truing data and 115 testing data for the total alkalinity and on page 1505 lines 1-2 they report 381 training data and 45 testing data. I struggle to understand how these numbers

add up? Are the distributions between training and testing data different for alkalinity and total inorganic carbon?

The training dataset is the same for total alkalinity and total dissolved inorganic carbon. This was done by choosing the stations were both parameters were simultaneously measured. However the validation dataset is different for A_T and C_T. The validation dataset for both parameters include the 10^{th} subset of the cross validation, but for A_T we also include the stations were the latter was measured without accompanying C_T. As for the numbers, we rechecked our dataset and corrected them accordingly. This section was rewritten as follows:

The dataset consists of 490 and 400 data points for A_T and C_T, respectively (Table 1). To ensure the same spatial and temporal coverage of the polynomial fits, the same training dataset was retained for both A_T and C_T. This was performed by selecting stations were both parameters were simultaneously measured; yielding 360 data points (Figure 1). To validate the general use of the proposed parameterizations we tested the algorithms with measurements which are not included in the fits (Validation dataset). For A_T, the validation dataset consists of 130 data points which are formed from the testing subset of the 10^{th} fold (40 data points), and from cruises where A_T was measured without accompanying C_T (90 data points). For C_T, the validation dataset is the same as the testing subset of the 10^{th} fold (40 data points).

The authors report that the algorithm was applied for polynomials of 1-3 (page 1504 line 19), however the authors do not explain why? Would it not be possible that a 4th order polynomial could further improve the total inorganic carbon fit?

We explain why a 4^{th} order polynomial could not improve the total inorganic carbon fit. We add in the text the following explanation:

High-order polynomials (4 and above) were discarded because they can be oscillatory between the data points, leading to a poorer fit to the data.

The authors use the established relationships to estimate alkalinity and DIC where there are no data, hence it is important to show that the algorithm does not overfit the data but is capable of extrapolating data, which is currently only partly done. E.g. one sign of overfitting would be if there is a substantial difference between the RMSE and mean difference between the residuals of the training set compared to the testing set. A table would help to illustrate this.

As recommended by the referee we tested the mean difference between the RMSE and mean residuals between the training set compared to the testing set. We add to the manuscript the following analysis:

- *For A_T we added in section 3.1:*

Furthermore, to make sure that the A_T algorithm does not overfit the data, we tested the difference in means between the RMSE and residuals between the training set compared to the testing set. The results show that for both RMSE and mean residual, we cannot reject the null hypothesis (that assumes equals means) between the training and validation datasets (Table 2).

Table 2. Mean difference t-test for the A_T algorithm between the training and validation datasets

	Training dataset	**Validation dataset**	
RMSE (µmol.kg^{-1})	**10.60**	**10.34**	**Mean difference t-test: H = 0; p = 0.83**
Mean residuals (µmol.kg^{-1})	**2.64e-13 ± 10.57**	**0.91 ± 10.30**	**Mean difference t-test: H = 0; p = 0.42**

- *For C_T we added in section 3.2:*

To make sure that the C_T algorithm does not overfit the data, we conducted the same analysis performed on the A_T datasets. The results show that for both RMSE and mean residual, we cannot reject the null hypothesis (that assumes equals means) between the training and validation datasets (Table 4).

Table 4. Mean difference t-test for the C_T algorithm between the training and validation datasets

	Training dataset	**Validation dataset**	
RMSE (µmol.kg^{-1})	**14.3**	**16.2**	**Mean difference t-test: H = 0; p = 0.04**
Mean residual (µmol.kg^{-1})	**-1.5e-12 ± 14.2**	**4.5 ± 17**	**Mean difference t-test: H = 0; p = 0.06**

Furthermore, it is somehow worrisome that the different algorithms from table 3 lead to such different results, as they are all developed for different regions, but do not seem to have a good predictable power in the Mediterranean.

The different algorithms presented in Table 3 are all developed in the Mediterranean Sea, except that of Lee et al. (2008). The reason why they lead to such different results is because they were developed over a limited time period, a limited geographical area, and with a limited number of data points. For instance, the Schneider el al. (2007) relationship is developed from only 15 data

points and during the months of October-November 2001. These relationships will hence tend to overfit our data and thus lead to such different results.

Minor comments:

I was very confused to see a reference to equation 1 in the text but I could not find the equation in the text, but rather had to look for it in table 1. It would help the reader if you could put equations in the text

We deleted table 2 and 4, and added instead Equation 1 and 2 in the text. Equation (1) was represented according to:

$$\mathbf{A_T = 2558.4 + 49.83(S-38.2) - 3.89(T-18) - 3.12(S-38.2)^2 - 1.06(T-18)^2} \quad \mathbf{(1)}$$

Valid for T > 13 ºC and 36.30 < S < 39.65

$$\mathbf{n = 360;\ r^2 = 0.96;\ RMSE = 10.6\ \mu mol.kg^{-1}}$$

Equation (2) was represented according to:

$$\mathbf{C_T = 2234 + 38.15(S-38.2) -14.38(T-17.7) - 4.48(S-38.2)^2 - 1.43(S-38.2)(T-17.7) + 9.62(T-17.7)^2 - 1.10(S-38.2)^3 + 3.53(T-17.7)(S-38.2)^2 + 1.47(S-38.2)(T-17.7)^2 - 4.61(T-17.7)^3} \quad \mathbf{(2)}$$

Valid for T > 13 ºC and 36.30 < S < 39.65

$$\mathbf{n = 360,\ r^2 = 0.90;\ RMSE = 14.3\ \mu mol.kg^{-1}}$$

Please clarify what you mean by summer and winter? E.g: is summer the average of the months of June, July and August?

This was added in the methods sections: '2.3. Climatological and seasonal mapping of A_T and C_T', as follows:

The summer seasonality is defined as the average of the months of July, August and September. The winter seasonality is defined as the average of the months of January, February, and March.

On page 1507 line 13 the authors mention the effect of biology; however, biology is not included in the polynomial fit. Why? You could e.g. use satellite derived biological proxies.

We mention the effect of biology only in reference to other studies such as: Bakker et al., 1999; Bates et al., 2006; Koffi et al., 2010; Lee et al., 2000; Sasse et al., 2013. The purpose is to mention that the parameterization of C_T is not only restrained to physical parameters. Also the aim of this paper is to derive A_T and C_T relationships from measurements of in situ parameters

such as temperature and salinity. This is why we did not include satellite derived biological proxies because it is out of the scope of this study.

Page 1507 line 6: ". . . presents a significant improvement . . ." please provide some information on how the significance has been tested.

This information was added.

In Equation 1, T and S contribute to 96% of the A_T variability and the RMSE of ± 10.6 µmol.kg^{-1} presents a significant improvement of the spatial and temporal estimations of A_T in the Mediterranean Sea surface waters (Mean difference t-test, H = 1; p = 0.04).