The authors thank the referee for taking the time to read through our manuscript. We are happy that a number of suggestions have been made to help improve the manuscript.

NOTE: The original comments by the referee have been numbered 1-15, and red text has been used for the response by the authors.

1. *One of the major concerns I have with this paper, is that the author's main aim appears to be to reduce the variability of the glider samples to match the significantly lower resolution CTD samples. The much higher temporal resolution and greater sampling area of the glider will give greater variability in the pHg compared to the pHCTD. Therefore, I am concerned that the authors may be misguided in their application of corrections – perhaps the difference in resolution could be commented on and the corrections discussed further, or the data presented in such a way that the pHCTD measurements are used as a guide rather than an elimination benchmark. This is discussed briefly in section 3.5 of Bresnahan et al., 2014. I understand that this correction of the sensor is based on the similarity of observed temperature and salinity measurements between CTD and glider – however, measurement techniques for these parameters are well established, with similar accuracy levels, and care should be taken when using the same standards for the ISFET pH sensor and pH calculated from bottled samples.*

We agree with the referee that variability seen by the glider could be different to the Ship measurements. We did see a larger extent of variability in the higher resolution profiles from the glider, compared with ship measurements. However, by comparing our ISFET measurements with observations by ship in the literature (as well as the ship measurements in this study), the extent in which pH should be expected to range in the northwestern Mediterranean Sea on a similar timescale is significantly smaller. For example, pH measurements (T = 25˚C) presented by Alvarez *et al*., (2014) taken from a hydrographic transect in the western Mediterranean Sea over a similar timescale to our deployment, varied between 7.87 and 7.93 at depths greater than 100 m, whereas pH measured by the glider's ISFET sensor (Fig. 4d) ranged between 7.97 and 8.18, which is roughly three times larger. This suggested that the measured range of the ISFET pH was incorrect (i.e. drifted) and required correcting.

A paragraph will be added to the introduction describing past observations of *in situ* high resolution pH in general (e.g. Hofmann *et al*., 2011), and hydrographic measurements specifically from the northwest Mediterranean Sea (e.g. Alvarez *et al.*, 2014) in order to provide readers a background of typical pH variability. We will make sure to comment on the difference in resolution in Sect. 3.

We agree with the referee that care should be taken when comparing temperature, salinity, and pH between the glider and the ship. However, we think this comparison is valid, as like with the sensors used to measure temperature and salinity, the accuracy of the Marianda VINDTA 3C and the ISFET pH sensor should have been similar (±0.005 ISFET pH e.g. Shitashima et al., 2002, and Marianda VINDTA 3C instrument determined by CRMs, and precision of 0.003 pH). Furthermore, the sparser ship measurements of temperature and salinity were within standard deviations of the high resolution glider measurements. From this, we can then assume that water mass properties were similarly represented in both ship and glider measurements, and that similarly this should be expected with pH.

We are aware that comparing ship and glider measurements of temperature and salinity represents physical properties only, so we will also compare dissolved oxygen concentrations observed by the glider

and by the ship.

1.1 The difference in variability could also be addressed with more information in the introduction on expected regional pH variability as seen from previous work in the Mediterranean (as briefly mentioned on page 5 line 14). This would demonstrate that temporal variability over the length of the deployment is minimal. Therefore, the procedures in the manuscript – correcting the data using 16 of the glider profiles, along with the pH of the bottled reference samples collected before the ISFET deployment time are valid for quality controlling the sensor.

As mentioned in the response to comment 1, the expected pH variability will be described in the updated manuscript. Past observations suggest that the extent to which ISFET pH varied in this trial over a similar timescale was greater, and therefore suggests the foundation in which we based our corrections is valid.

2. *P3 Section 2.2: More information on the ISFET-sensor used would be useful – specifically the calibration.*

More information of the ISFET sensor will be included in the manuscript, relating to the sensor itself and the calibration procedure. Some details that will be added to the manuscript can be found in the responses to the proceeding 9 comments, and in review 2 – comments 9.1-9.5.

2.1 It would also be interesting to know what the authors mean by poor quality –was this caused by integration into the glider electronics, or did the sensors malfunction? A brief sentence on this would also be useful – given that the paper is based around discussing challenges when field-testing sensors.

As discussed in Review 2 – comment 11, we think the regular on/off cycling of electricity to the integrated dual sensor in between sampling did not allow it to function properly. A few sentences explaining this will be added to the updated manuscript.

2.2 The authors specify that they used a Cl-ISE. How long was this conditioned for? Previous studies (Bresnahan et al., 2014, Takeshita et al., 2014) both recommended conditioning in seawater levels of bromide ions before deployment to prevent reference electrode drifts.

The ISFET and Cl-ISE were stored in a bucket of seawater for an hour before the deployment of the glider. The salinity of this water was about 38.05. This information will be added to the manuscript.

2.3 What was the ionic strength of the two buffers used on deck to calibrate the ISFET?

The buffers were made up in synthetic seawater of S = 35, which would have an ionic strength of about 0.7 M. A sentence about this will be added to the manuscript.

2.4 You also specify the pH of these solutions to a 4 decimal point (5 sig. figs). This is very accurate for a pH sensor – particularly when the accuracy of the pH sensor you deploy is only 0.005. What pH system did you use to get this accurate buffer pH to calibrate your solutions?

The buffer solutions of AMP and TRIS were created following SOP 6 from Dickson et al., (2007) and the pH values of these buffer solutions were taken from this reference, assuming that the temperature of the

solutions was 25˚C, and the salinity = 35. We did not measure the pH of the solutions by any other means. We will reduce the decimal points of the pH values listed in the manuscript, and we will add this information about the buffer solutions to the manuscript.

2.5 Was the deployed ISFET-measured pH of the buffer solutions the same before and after (i.e. was there any drift?)? Were the same solutions used – was there any drift in the solutions?

The same buffer solution batches were used before and after the deployment. The ISFET measured values of the buffer solutions at the end of the deployment differed to those measured before the deployment. This drift was corrected for using the calibration data before and after the deployment. This information will be added to the updated manuscript.

2.6 Was there any noticeable biofouling on the ISFET sensor during the deployment?

It was clear after an inspection of the glider and sensor that there was no biofouling. We will state this in the manuscript.

2.7 Was there any lab-based temperature calibration done prior to deployment? Bresnahan et al., 2014 discuss a temperature error of <0.015 in their calibration of the sensors – this is greater than the specified accuracy of the deployed ISFET sensors.

The ISFET sensor was supposed to take into account temperature and pressure changes in the environment (Shitashima *et al*., 2002; Shitashima, 2010), hence no lab-based temperature calibration was performed.

2.8 You mention the air temperature when calibrating with the buffer solutions, a measurement of the temperature of the buffer solutions would also be useful, particularly as you later correct for temperature dependence of the sensor. This is important, as the temperature of the solution may change the buffer pH (particularly when using such accurate pH figures) between the pre-deployment measurement and post-deployment measurement.

We will list the recorded temperature ranges of the buffer solutions before and after the deployment in the manuscript and we will comment on the uncertainty relating to possible pH changes of the buffer solutions as a result of changing temperature.

2.9 Finally, you provide a reference to Fukuba et al., 2008. This particular ISFET sensor does not have details of correction using buffers before and after deployment, but rather buffer solutions deployed with the sensor itself, allowing for in situ referencing. This is not the same procedure as the sentence is suggesting, nor does it provide an example of the converting the raw output to pH. Unless the ISFET sensor deployed had a similar "self-calibration" system, I would suggest removing this reference.

This reference will be removed from the manuscript.

3. *P4 Line 18: the difference in the DIC and the TA quoted from replicate samples – is this calculated from the standard deviation for each replicate? You state, in the previous sentence, there were two to three replicates collected per CTD cast – If this is not the standard deviation, how was this difference calculated between the three samples.*

We calculated the mean absolute difference between replicate samples, with the value to the right of the '±' symbol representing the standard deviation of these absolute differences. However, we now think it will be better to list the mean standard deviation of the replicate samples in the updated manuscript.

4. *P4 Line 20: Please also state the borate-chlorinity ratio and the sulphate constants that were applied when using CO2SYS- with appropriate references. I realise these may be quoted in the best practices section in the paper by Orr et al (2015), however it would be best if they were also specified here for clear understanding.*

The suggested ratio and constants will be stated in the updated manuscript.

5. *P4 Line 32: I find the range of standard deviations quoted throughout the manuscript to be confusing. For each specified bin (top 150m and below 150m) there is range of standard deviations quoted instead of one number for each bin. Is the standard deviation not calculated over the whole 150m? Is it further subdivided into smaller bins, and in which case what size are these bins and how many are there? I feel this should be clarified at the start of this section as the ranges are applied throughout the remainder of the manuscript. I assume these bins are the same as those specified in the caption for figure 5, but should be mentioned in the text for clarity.*

To calculate these ranges of standard deviations, the values from all profiles of a given variable (e.g. DIC,A_T, pH_s…) were sorted into 10 m depth bins down to a maximum depth of 1000 m. The mean and standard deviation was calculated for each one of these 10 m bins using the assorted data within. This produced two arrays; 100 x mean values and 100 x corresponding standard deviation values between the surface and 1000 m depth. Thus, the quoted standard deviation ranges (e.g. for the top 150 m) were defined using the minimum and maximum standard deviation calculated from these bins within the depth range (e.g. 15 out of 100 binned standard deviations for the top 150 m). We will make sure to explain clearly how these standard deviation ranges were calculated in the updated manuscript before such ranges are listed.

6. *P5 Line3: The authors refer to environmental variability when referring to the range of pH observed. This is not further discussed - What is the expected natural variability for the region? How much extra variability was observed and can be attributed to instrumental error? I realise that this is mentioned briefly in line 12, however numbers specifying the expected pH range and variability would be useful for those of us with little knowledge of the region.*

As mentioned in comment 1 above, a paragraph will be added to the introduction describing past observations of *in situ* high resolution pH in general, and hydrographic pH measurements specifically from the Mediterranean Sea. A discussion of the comparison between the natural variability of past observations by ship and the ISFET measurements on a similar timescale of a few weeks will be added to the manuscript. As the pH derived from bottle samples in this deployment varied within a similar range to past observations of pH (roughly 0.1 pH) when considering all measurements between the surface and 1000 m depth, the difference between the standard deviations of glider measurements and ship measurements in this study could be used as an indication of the instrumental error. The instrumental

error would therefore have been between 0.03 and 0.09 in the top 150 m of the water column, and between 0.005 and 0.045 beneath this.

7. *Furthermore, the instrumental error is not discussed in section 2.3. I think the authors meant sections 3.2 and 3.3.*

We thank the referee for highlighting a possible mistake. However, the authors were referring to the instrumental error associated with obtaining *c*(DIC) and A_T samples using the VINDTA instrument, which would have in part contributed to the standard deviation values obtained from the pH_S measurements. This is described on page 4, lines 18-19 in section 2.3.

We will alter this sentence to make this clearer.

8. *P5 Line 22: Please specify if the same subtraction was performed on the salinity, dissolved oxygen and potential temperature.*

This will be specified in the updated version of the manuscript.

9. *P6 Line5: Does the ISFET have a constant offset caused by light? Or an offset changing with irradiance time/strength? Could you give some indication of the size of the offset based on your experiments.*

The offset depended on irradiance strength (i.e. the value changed depending on how close the sensor was from the light source, and the type of bulb used), and remained relatively constant when the light source was turned on. The offset was roughly between -4 and -6 x 10^8 counts and between -1 and -2 x10^8 counts when the LED and Halogen lights were used, respectively. A sentence will be added to this paragraph describing these observations.

10. *P6 Line 28: I find it confusing when you discuss a constant depth –time varying offset, and then subsequently refer to, what I assume is the same correction, as a constant offset. It is not a constant offset as it varies with time. It also presumably varies with depth, as the correction was determined from the depth where the potential temperature was 14°C.*

Offset values were derived using the difference between mean pHs and pHg where the temperature of the water was 14 ˚C, and the depth of this indeed varied throughout the time period of the deployment. However, the authors refer to the method in which the offset was applied. In other words, the offset applied to the glider data did not change with depth (now referred to as 'depth-constant', see Review 2 – comment 18), but changed in time (i.e. a different offset value was determined for each dive profile). This will be further clarified in the manuscript.

We also will not refer to the offset as just 'constant', as this is not correct as highlighted by the referee.

11. *P7 Line 9: It would be good if the authors could specify the slope and the intercept of the linear regression in the text. This will allow better comparison with other studies.*

More information will be added to the section. This will include the offset equation, delta pH equation, and temperature and pressure correction equations, incorporating the calculated slope and intercept coefficient values. These will be compared with other findings, such as Johnson *et al.*, 2016.

12. *P7 Line 27: The authors say poor-accuracy, is this relative to previous deployment? How did they determine the accuracy if the paper is based around correcting the pH sensor to the bottle samples? The best accuracy quotable for the sensor is that related to the reference samples.*

The authors were referring to the ship based reference samples when stating the poor-accuracy of the ISFET measurements. This will be clarified in the text.

13. *P8 Line 7: Remove "there being"*

This will be removed.

14. *Conclusions: The conclusion could be improved by summarising the findings of the paper including the biogeochemical variability (similar to the abstract). The authors also specify that the corrections they performed are not generally recommended or valid. A brief discussion of why these corrections are valid in this study, and under what other conditions they may not be valid would be good for future work by other studies.*

The conclusions section will be expanded to include findings on the physical and biogeochemical variability, and we plan to discuss the points made by the referee regarding the corrections.

15. *Figures: (in general) seem to have a grey line around the edges. This is particularly on figure 8 where it looks like another figure was cropped out.*

This was caused when editing the plots and will be removed. The plots will now also be uploaded as PDFs for better quality.