Response to the comments of Reviewer-1

Thanks for the detailed comments. We are providing detailed response to each comment.

The stated purpose is: ". . . on understanding the process controls in estimating winter nutrient loadings. . .", but the majority of the manuscript is devoted to presenting and validating the PCR and CCA models (4 out of 6 tables and 5 out of 8 figures). From this, it seems like the purpose of the paper is to demonstrate (and compare?) two models that can incorporate precipitation forecasts to provide nutrient forecasts. This is important because despite forecast improvements, they are currently underutilized, and there aren't many examples of how they can be used for water quality forecasting. In addition, to better highlight the main points and conclusions, the results and descriptions need to be sharpened and focused (specific examples follow in Specific comments).

The purpose of the manuscript is to identify the process controls through model development as well as through diagnostic analysis. Key processes that control TN loadings and provide useful information in improving season-ahead predictions are: (a) precipitation (b) basin storage represented by previous-month streamflow and regionally the process that influence the nutrient variability is El-Nino Southern Oscillation.

Sentence in lines 78-80 reflects this: The purpose is to understand the *"controls"* that are required for developing a skillful seasonal nutrient forecasts and also to assess how the skill in hydroclimatic predictions translate to skill in nutrient forecasts over the regional scale.

One main technical question arose: how are the climate forecast errors considered? You consider R2LOADEST and R2CCA/PCR, but do these consider the possibility that the precipitation forecast for JFM is "wet" but it turns out to be "dry"? (For more details and specific places in the manuscript related to this comment, see Specific Comments).

	Observation		
Station Index	**BN**	**N**	**AN**
2	(0.57,0)	(0.57,0)	(0,0.14)
3	(0.29,0)	(0.29,0.10)	(0,0.29)
4	(0.29,0.05)	(0.43,0)	(0,0.14)
7	(0.29,0.05)	(0.43,0.10)	(0,0.43)
8	(0.57,0)	(0.43,0.14)	(0.14,0.29)
9	(0.43,0.05)	(0.43,0.10)	(0.14,0.29)
10	(0.14,0.1)	(0.43,0.05)	(0,0.43)
14	(0.14,0.05)	(0.14,0.14)	(0.14,0.43)
15	(0.16,0.05)	(0.33,0.05)	(0,0.33)
Average	(0.32,0.04)	(0.39,0.08)	(0.05,0.31)

This is a very good comment. We are presenting a table (above) that shows when the observed precipitation is below-normal (BN), Normal (N) and Above-Normal (AN), the probabilities of nutrient forecasts predicting a wrong category. For instance, the values under BN denote the probability of nutrient forecasts predicting normal and above normal (in the same order inside the parenthesis) events. Similarly values under normal (above-normal) denote the probability of nutrient forecasts predicting below-normal and above-normal (below-normal and normal) events. BN, N and AN for precipitation and nutrient forecasts are obtained if the observed or the predicted values in a given year fall within <33 percentile, 33-67 percentile and >67 percentiles of their respective climatology (i.e., based on observed/predicted values from 1987-2007) respectively. The table is summarized for the nutrient forecasts developed using precipitation forecasts and December streamflow as predictors under SSV (Table 6) for stations that exhibit significant skill in predicting nutrient forecasts.

Please note that this is different from the requested information, since we are comparing the performance of the nutrient forecasts with the observed precipitation, not with the precipitation forecasts. Hence, this is more stringent than what was asked.

From the above table, we infer that the probability of wrongly predicting a particular category is around 50%. For instance, in station 2, if the observed precipitation is BN, the forecasted nutrients fell into normal 57% of the time and above-normal 0% of the time (0.57, 0). However, the entire analysis is based on 21 years (1987-2007) of seasonal nutrient forecasts with 7 events being fallen into BN, Normal and Above-normal categories. This is really a very small sample size to draw definitive conclusions. However, if we average across all the stations, number of forecasts for each category increases (7*9). Based on that, we infer that the probability of predicting each category wrong varies between 0.36-0.40. This is quite reasonable given that we are using simulated nutrients.

We can include the above table as Table 7 in the revised manuscript and also include the above discussion.

Specific comments: Title: Consider removing second "variability" (perhaps replace with "concentrations" or "loadings"). Consider revising to better reflect purpose (see subsequent comments).

I think the variability on nutrients is important, since we focus on the interannual variability of nutrient loadings as well. However, to reflect the importance of loadings, we can revise the title as: Interannual Hydroclimatic Variability and its Influence on Winter Nutrients Loadings Variability over the Southeast United States

Abstract: Define abbreviations, remove "Table 2", and revise sentence: "Stations that have very high R2(LOADEST) (>0.8) in predicting the observed WQN loadings during the winter (Table 2) exhibit significant skill in loadings."

We agree. The sentence could be revised as: *S*tations that have very high coefficient of determination (> 0.8) in predicting the observed WQN loadings during JFM exhibit significant skill in predicting seasonal TN loadings using climate forecasts.

Introduction: This does a good job of motivating the study, but could use a revision for sentence and word-smithing to insure that there is logical flow and natural transitions between ideas.

p. 10937 line 11. Remove sentence or edit: "Thus, it is critical to estimate the seasonal nutrient loadings conditioned on the expected runoff from nonpoint sources."

The sentence could be revised as follows: Thus, for these virgin basins, it is critical to estimate the seasonal nutrient loadings due to the potential changes in hydroclimate during the season.

p. 10937 line 16. Define ENSO, SST in main text.

Thanks. We can incorporate this in the revised submission.

p. 10937 – In this section, consider revising to make a more explicit connection between SSTs, ENSO, and the precipitation forecasts that you will be using to make nutrient forecasts: SSTs drive ENSO, which drives precip/streamflow patterns, which drives nutrient concentrations.

I think the entire paragraph reflects it. Further, the association between SST variability to nutrients variability is inferred from the study through rigorous analyses, which we plan to move before Section 3.1 (see the response to the comments on organizing the manuscript). Hence, we don't want to write as if this has been investigated or known. But, the paragraph sets up the tone for the linking SST and nutrients. So, we would like it to leave it as such.

p. 10938 lines 18-21. Clarification is needed to explain how the purpose of the paper is to "understand the "controls" that are required: : :". The paper does not seem to explore "controls", i.e., controlling processes or mechanisms driving nutrient concentrations. That might come into play if you were testing a suite of predictors to find the "best" combination for nutrient forecasting, but that does not seem to be the focus (you look at precip, and then briefly at flow and ENSO). It seems to be demonstrating statistical tools that one can use to incorporate seasonal forecasts to develop nutrient forecasts. Or to compare different statistical tools to see what types of models are best suited to incorporate seasonal forecasts for nutrient forecasting.

Key processes that control TN loadings and provide useful information in improving season-ahead predictions are: (a) precipitation (b) basin storage represented by previous-month streamflow and regionally the process that influence the nutrient variability is El-Nino Southern Oscillation. The purpose of the manuscript is to identify the process controls through model development as well as through diagnostic analysis. For instance, Figures 1 and 7 demonstrate it just using correlation (no model development), whereas Figure 6 shows it through model development. Previous month streamflow basically indicates the importance of basin storage. Understanding the influence of the predictor through model development is certainly more rigorous since all the model results are presented through validation. That is why we emphasize the importance of "controls" in developing skillful forecasts. Hope we explained it here.

p. 10938 lines 21-24. Replace "climate forecasts" with "precipitation forecasts". Replace "land surface conditions" with "flow conditions".

We agree. We can change the sentence accordingly.

Section 2. Data Sources. Consider reorganizing to clarify that there are 2-steps in this section. For instance: First, streamflow is used with the LOADEST to simulate a full-record of water quality; then it is used as a predictor for season-ahead nutrient forecasting. One idea is making Section 2. "Study Area and Data". You could start with the study area description, then introduce the data sources in light of the two steps: (i) Water quality simulation and (ii) Season-ahead forecast predictors.

We think the section reads fine. Just for clarity, we can add couple of sentences to describe the study area before introducing streamflow data. We are presenting it below.

2. Study Area and Data Sources

We consider 18 watersheds (Figure 1) from the Southeast US for understanding the role of hydroclimate in influencing interannual nutrient variability. Previous studies have shown that winter precipitation and streamflow over the Southeast US are heavily influenced by the ENSO variations (Ropelewski and Halpert 1987, Devineni and Sankarasubramanian 2010). The selected 18 watersheds span over seven states and the streamflow with drainage area ranging from 136 km^2 to 44547 km^2 (Table 1). Given that the selected watersheds are minimally impacted by anthropogenic influences, we hypothesize that the interannual variability in winter nutrients could be explained by the precipitation variability as well as by the antecedent flow conditions. For this purpose, we assemble hydroclimatic and water quality databases for developing season-ahead nutrient forecasts over these 18 watersheds.

Section 2.3. Briefly, what kind of model is LOADEST? E.g., Mechanistic, empirical, other?

LOADEST model is a statistical model that estimates daily loadings based on the observed daily streamflow and the centered time (dtime) of the year of the observation (Runkel et al. 2004). We also feel it is appropriate to include the basic regression form (given below) that is used in the LOADEST model as equation 1, so that it helps the reader to understand Table 2.

$$\ln(\hat{L}) = a0 \; + a1 \ln(Q) + a2 \ln Q^2 + a3 \sin(2\pi dtime) + a4 \cos(2\pi dtime) \;\; \ldots(1)$$

Dtime is the centered decimal time in years as defined in Cohn (1992) and in Runkel et al., (2005).

p. 10940 lines 24-27: Consider revising this to be less technical and more descriptive. For instance, consider removing "dtime" and instead describe why it was appropriate to exclude a time trend in the regression. Further, you indicate what's not included (i.e., the time trend), but is there a concise way to summarize what predictors are included? You could describe them instead of listing the model numbers (i.e., 1, 2, 4 and 6). Listing the predictors gives a sense of what the important parameters are for simulating the water quality data.

I think including the above equation will address this question.

Table 2. Consider revising the table to only include the most necessary information, e.g., may not need station number and model number. The coefficients values may not be necessary, especially if we don't know what each predictor is. Or could you summarize the results for most of the models, but only show results from 2 locations, say the best fit and the worst fit. E.g. station 17 vs. station 18 (?).

We will remove the station number and the model number. Since we are going to provide the basic equation form for the LOADEST model, it makes sense to include the coefficients.

p. 10941 line 5. Replace "in predicting" with "of". The reason being that "goodness-of fit" only implies a good fit to the data, it doesn't test the model in a predictive mode.

We agree.

p. 10941 lines 6-23. Revise paragraph to better highlight the key points and procedures. Point out the key results for the JFM LOADEST simulation from Table 2. Much later on p. 10953 you note: "stations that have very high R2 (LOADEST) (>0.8) in predicting the observed WQN loadings during the winter (Table 2) exhibit significant skill in loadings." In this paragraph, you should point out the stations with low R2 values that you refer to later (e.g., station 5, 6, 18: : :).

We can add two sentences in line 16: From Table 2, we clearly see that the performance of the LOADEST model in predicting JFM nutrients is poor in stations 5, 6, 13, 16 and 18.

Section 2.4. Technical comment: How accurate/reliable are the 3-month ahead precipitation forecasts that you use? That is, let's say the retrospective forecast for JFM of 1989 was wet, is it possible that it actually ended up being dry? This would contribute to errors in your results. Consider revising the technical description of how the forecasts are constructed (e.g., "To force the ECHAM4.5 with SST forecasts, retrospective monthly SST forecasts were developed based on the observed SST conditions in that month based on the constructed analogue approach") to better highlight how that method affected the confidence/correctness of the forecasts that you use. (E.g., consider: "By forcing the ECHAM4.5 with an SST approach based on xyz, it insured that the precip forecasts were lmnop...")

This relates to the technical comment discussed in the overview of the comments. Please look at the table in the first page of this response and the associated response/discussion.

With regard to revising the precipitation forecasts development, we feel it could stay as such. Seasonal climate forecasts are typically developed either using Atmospheric GCMs (AGCMs) or using Coupled GCMs (CGCMs). In the case of former, it is a two-tiered system, in which SSTs are forecasted first using a statistical/dynamical model and then they are forced with AGCMs. In CGCMs, since ocean and atmospheric models are coupled, they are run in a continuous mode. We can add the following sentences to substantiate the precipitation forecasts skill.

Seasonal climate forecasts are typically developed either using Atmospheric GCMs (AGCMs) or using Coupled GCMs (CGCMs). In the case of former, it is a two-tiered system, in which SSTs are forecasted first using a statistical/dynamical model and then they are forced with AGCMs. In CGCMs, since ocean and atmospheric models are coupled, they are run in a continuous mode. Recent studies clearly show that AGCMs are more skillful than CGCMs (Goddard et al. 2003). Further, Devineni and Sankarasubramanian (2010) show that ECHAM4.5 precipitation forecasts explain 25-36% of the variability in observed precipitation over the Southeast US. For this study, we utilize the retrospective winter precipitation forecasts from ECHAM4.5 General circulation model forced with (http://iridl.ldeo.columbia.edu/SOURCES/.IRI/.FD/.ECHAM4p5/.Forecast/ca_sst/.ensemble24/.MONTHLY/.prec/, International Research Institute of Climate and Society (IRI) data library) (Li and Goddard 2005) constructed analogue SSTs.

Figure 2. Good figure.

Thanks.

Table 3. It's very interesting that there is little to no difference between the variance explained by PC1 of Q or TN. This helps to justify what you are doing (i.e., going directly from precip to TN, by-passing concurrent streamflow). I think this could be highlighted in the text more. (See 10943 line 10-12).

We can add the information below to substantiate this further:

Given that the PC1 of precipitation forecasts explain almost the same amount of observed variability in precipitation, streamflow and TN (Table 3 and Figure 2), it is logical to develop TN forecasts conditioned on precipitation forecasts. This is expected since the selected basins are virgin watersheds with non-point loadings being the primary source of TN.

Figure 1 caption: Add "precipitation forecast" before "grid points".

We agree. This could be revised.

Section 3. Technical comment: From the paper you indicate: ": : : we first identify relevant grid points (Table 3) of JFM precipitation forecasts that have statistically significant correlation with JFM observed precipitation for each watershed. Nearest grid points that are significantly correlated to each watershed (Fig. 1) are selected." Is this how you insure that the forecasts are accurate/reliable? It might help to add information on why you did this here.

Given that the precipitation forecasts have spatial bias, it is important to identify the relevant grid points that explain the variability in the observed precipitation, streamflow and TN at the watershed scale. Hence, we identify the grid points based on correlation between observed precipitation and forecasted precipitation. Thus, identifying the relevant grid points whose forecasted precipitation associate well with basin hydroclimate help us to reduce the spatial bias and recalibrate the precipitation forecasts using principal components.

In the Discussion you later say: "By selecting grid points of precipitation forecasts that are statistically significant with the observed precipitation in the basin, we ensure that the skill in predicting nutrient loadings is related to the basin process as well." I'm unsure of what you mean by "related to the basin process". If this step is not to insure that the forecasts are accurate/reliable, then you would have to add the forecast error too (i.e., Independent errors would be from (i) LOADEST simulation, (ii) low-dimensional model, (iii) forecast model).

We imply basin process here as the observed precipitation over the basin. This is primarily carried to make sure that the forecasts are reliable and explain the variability in observed precipitation.

Section 3.1. It would be helpful if you could point out the difference/advantage of developing these low-dimensional models as compared to just using the regressions already developed from LOADEST. This might become clear when you add information about the LOADEST predictors. The low-dimensional models are suited to using the GCM forecasts, but do any of the LOADEST models contain precipitation, or other large-scale information, as predictors?

The LOADEST models estimate daily loadings from observed streamflow and the time of observation by developing regression relationships between streamflow and loadings. So, it is not appropriate to use LOADEST models to predict loadings directly from precipitation, since LOADEST model requires daily streamflow, whereas the forecasted precipitation is available at monthly to seasonal time scale.

p. 10943, line 21. Consider revising this sentence, not sure what is meant by "marginal bias".

Marginal bias implies bias in estimating the long-term mean of the predictand, which is winter TN in this case. Thus, PCR is also a recalibration approach that eliminates the bias in estimating the long-term mean.

Section 3.1. It might help here to clarify the paper purpose here. (See earlier comments).

We explained this early. The purpose here is in identifying predictors or process controls that play an important role in developing skill forecasts. We don't want to consider observed variables during the season to develop TN forecasts for the season. Hence, we consider predictors that are available at the beginning of the season (e.g., December streamflow) or the predictors for which the forecasted values are available (e.g., forecasted precipitation from ECHAM4.5) during the season. Apart from these two variables, we also identify ENSO as a predictor that influence at the regional scale. We plan on moving Section 4.4 on ENSO before section 3.1.

Table 5, Figure 3, Figure 4. This table and two figures show the validation for PCR models in terms of RMSE and R2 from LCV and SSV. I'm wondering if there is a way to distill these results to show the key conclusions. For instance, with PCR the stated conclusion is: "Thus, based on two different validation methods, we understand that eleven stations (2–4, 7–11 and 13–15) exhibit statistically significant skill in predicting the observed WQN loadings using the PCR model developed separately for each site." These are conclusions based mainly on the validation method that had the most conservative result – namely Figure 4 (it found less sites significant than did Figure 3). It might be interesting to only show the results *from Figure 3 that are different than what you found in Figure 4 (e.g., site 17 is significant in Figure 3 but not in Figure 4, why?). What are the most important values in Table 5 RMSE, i.e., which parts bolster your conclusions?*

We are not sure how we can summarize this better. It will only lead to loss of information. Further, if we are going to show information in Figure 4 which is different from Figure 3 (i.e., by removing stations), it is fine in showing the information for all the stations.

Reviewer is asking an interesting question on why site 17 shows significant results under Figure 3, but not under Figure 4. LCV exploits all the available data in the future years to predict the TN loadings for the left-out year. Thus, LCV shows that there is potential in developing seasonal nutrient forecasts for site 17. However, under SSV, we cannot guarantee that skill in developing real time (without using future information) forecasts, since the trained model using the data from 1957- 1986 is not capable of developing a statistically significant forecast for the period 1987-2007. But, as we collect more data in the future, we may be able to develop statistically significant forecasts for station 17. Thus, LCV shows the potential skill in developing the forecast, whereas SSV shows the demonstrable skill in developing real-time forecasts. Thus, in our summary and conclusion, we list the stations that perform well under LCV and SSV. We can include the above discussion in the revised manuscript.

This way of summarizing the results may be conservatory. But, we feel that is better to do that way, since the simulated nutrients used for developing the PCR model is estimated using the LOADEST model based on observed streamflow . This is another reason why we account the errors in the LOADEST model too in reporting results using equations (2) and (3).

Table 5, Figure 5, Figure 6. This table and two figures show the validation for CCA models. Similar to above comments, can these be distilled? Would it make sense to only show the more conservative results – i.e., where less sites are significant (Figure 6), and then only the stations that show different results from Figure 5?

Same response to the previous question applies here. It is better to present the results as such. The reason we consider CCA as an additional low-dimensional model primarily stems from Figure 8. Given that we infer that ENSO influences the winter hydroclimate over the Southeast US, can we develop regional statistical models for developing nutrient forecasts. Based on this approach, we develop only three regression models by exploiting the spatial correlation among the variables. However, results show developing models for individual sites using PCR performs better. We can include this discussion in the text on why we consider CCA as an alternate low-dimensional model.

Also, why do you show the result from the LCV for CCA as a map (Figure 5) versus the LCV for PCR with box plots (Figure 3)? You might want to comment on why you think the SSV results in less significant sites than the LCV.

There is no specific reason. We feel that maps show the skill in a spatial sense. For instance, most of the coastal watersheds in Florida perform better. In fact, we can change Figure 3 to a map by plotting the median value of the R^2.

p. 10947 line 7. Replace "However" with "Here".

We can revise this.

p. 10949 line 25. Replace "CCA model in explaining" with "the CCA model to explain"

Agree. This could be revised.

Section 4.3. This section has interesting information, but it comes as a bit of a surprise in the manuscript. Although it is mentioned in the Abstract, it is not mentioned in the Introduction, Section 3, or Section 4. Up to this point, the majority of the paper indicates that you are only going to look at precipitation forecasts as a predictor of water quality. Was this work motivated from the results of the low-dimensional models? Consider removing or revise so that this fits more logically. (Perhaps move this to the discussion section?) Or introduce it earlier and indicate that Section 4 includes results from two parts: (i) using ECHAM4.5 precipitation forecasts alone and (ii) adding antecedent streamflow.

Thanks for the comment. I think moving this to the discussion section would make more sense.

Section 4.4. Similar to section 4.3 – this section is interesting but does not fit where it is in the manuscript. Consider removing or revise so that this fits more logically. (One idea: This is more of a motivation – it could almost go after Figure 2. Fig 2 establishes that there is a strong precip-TN correlation, but this would be going one step further by saying there is also a strong correlation between ENSO and TN. Further, it would be good because precip forecasts tend to have more skill during ENSO years).

Nice comment in organizing the manuscript. We can move the entire section 4.3 before Section 3.1.

Section 5. Consider removing some information in the discussion that is more of a summary than discussion (or move to a summary or methods section): e.g., "Since obtaining long continuous records of daily observations of nutrients is difficult particularly over a large region, we employed simulated nutrient loadings from the LOADEST model to understand the role of climate variability in modulating the interannual variability in nutrients over the SEUS. However, to account for the errors in the LOADEST model in predicting the observed WQN database, the reported skill measures (Eqs. 2 and 3), R2 and RMSE, are adjusted for both LOADEST model error as well as the error of the low-dimensional models."

Agree. We can remove this, since it is more of a summary. We intend to bring Section 4.2 inside the discussion section.

I'm uncertain about the statement made here in the Discussion: "Thus, the intent of this study is to understand how well climate and basin storage conditions control the seasonal TN loadings rather than developing a skillful nutrient forecasts using lowdimensional models." The majority of the paper is devoted to evaluating the skill of nutrient forecasts using low-dimensional models, so I'm uncertain what point this is trying to get across. This goes back to the question of the purpose of the paper. I interpret the intent of the study to demonstrate the potential of climate forecasts for water quality prediction. The results imply that while these models may not be ready for operational forecasting, there is great potential to further develop water quality forecasting tools.

I agree. It was not written very clearly. We can revise it as follows: Thus, the intent of this study is to understand how well climate and basin storage conditions control the development of skillful forecasts of TN loadings.

The paragraph that begins: "Perhaps the most important utility: : :" makes a good case for the utility of the season-ahead forecasts of nutrient loadings. This might be a good place to discuss the comparison between the results of the CCA and the PCR. For instance, if a manager wanted to use one of your models for water quality trading, which model – the CCA or the PCR or a hybrid – and for what cases would you recommend each? What are the advantages/disadvantages of each, which had better results for what? It would strengthen your paper if you could link these applications more closely to the analysis and results of this paper – even if it is only hypothetical. This might also be a good opportunity to motivate the results from section 4.3. You could indicate that to implement in practice, you need to improve the forecasts, which led to including the antecedent conditions.

I think we would like to leave the CCA and PCR results in Section 4.1. Given that we are moving section 4.2 to discussion and section 4.3 before section 3.1, the flow of the manuscript should be good.

On using these results for management, the performance of the PCR models developed using precipitation forecasts and streamflow is better. However, we don't want to overemphasize the management since all our results are based on virgin basins. Our results show that there is potential in developing nutrient forecasts which could be utilized in developing nutrient management strategies. We are currently preparing a manuscript that uses weather forecasts and climate forecasts to develop an integrated strategy for managing both point and non-point loadings.

References:

Cohn, T. A., Caulder, D. L., Gilroy, E. J., Zynjuk, L. D., and Summers, R. M.: The Validity of a Simple Statistical-Model for Estimating Fluvial Constituent Loads - an Empirical-Study Involving Nutrient Loads Entering Chesapeake Bay, Water Resour. Res., 28, 2353-2363, 1992.

Devineni, N., and Sankarasubramanian, A.: Improved categorical winter precipitation forecasts through multimodel combinations of coupled GCMs, Geophys. Res. Lett., 37, L24704, 2010.

Goddard, L., A. G. Barnston, and S. J. Mason, 2003: Evaluation of the IRI's ''net assessment'' seasonal climate forecasts: 1997–2001.Bull. Amer. Meteor. Soc., 84, 1761–1781.

Li, S., and Goddard, L.: Retrospective Forecasts with the ECHAM4.5 AGCM, IRI Technical Report, 05-02, 2005.

Runkel, R, L., Crawford, C. G., and Cohn, T.A.: Load Estimator (LOADEST): A FORTRAN Program for Estimating Constituent Loads in Streams and Rivers, U.S. Geological Survey Report, 2004.

Ropelewski, C. F., and Halpert, M. S.: Global and Regional Scale Precipitation Patterns Associated with the El-Nino Southern Oscillation, Mon. Weather Rev., 115(8), 1606-1626, 1987.