We thank the second referee for valuable comments on our manuscript. All comments are summarized with a numbering style and corresponding responses were stated after an arrow symbol (➔). The line numbers (Line) referenced will be changed for the final version of the revised manuscript.

<u>Reviewer #2</u>

General comments:

2.1. **1)** First, given that changes such as increase in CO2, temperature and precipitation are likely to occur simultaneously, what is the rationale for assessing the effects separately? This seems particularly tricky given that often temperature and precipitation can have opposite impacts on streamflow and nitrate loads. More justification of this choice would be helpful, as well as some discussion on how separating these changes might impact the results of the paper. **2)** The GCMs do include multiple changes simultaneously, but because the change in precipitation and temperature in the GCM runs are different from those in the "sensitivity runs", it is difficult to understand the impacts of the difference in changes versus the consideration of simultaneous changes in multiple factors. For example, it would helpful to know if there an increase in temperature that would cancel out simultaneous increases in precipitation?

➔ **1)** This paper analyzed the climate change impacts on crop growth and related nitrogen cycling/transport processes in an agricultural catchment, the typical representative of the coastal watershed in the CBW. The climate change impacts were represented by two steps (sensitivity and GCM scenarios). The first step was to investigate the individual effects of the key climate factors on the crop biomass, water and nitrate cycling. This step was to develop in-depth knowledge and understanding on how each climate factor affects these underlying processes (Wolock and McCabe, 1999). The second experiment (e.g., simulation with the GCM output) was to quantify and predict these crop effects on water and nitrogen cycling at the local catchment level, with respect to the foreseeable climate changes. We used the GCM projections to describe foreseeable changes, as the combination of climate factors and their interactions could not provide complete climate change/variability information (including seasonal and inter-decadal variability, Mearns, 2001). For example, crop growth and agricultural nutrient loadings (e.g., fertilizer-driven nutrients) are highly sensitive to inter-monthly variations of the climate system. However, such variations in climate system could not be captured by a combination of three climate sensitivity scenarios. In our revision, we will clarify the purpose of the study design in the introduction as suggested by the reviewer.

➔ **2)** As requested by other reviewers, we replaced the existing GCM data with the state-of-the-art GCM data (CMIP5) and these data indicate increases in temperature and precipitation compared to the baseline scenario (see Figure 7). Thus, we did not consider the precipitation decrease sensitivity scenario because the climate pattern shown in new GCM data well matched with the sensitivity scenario.

2.2. Second, how general are these results–for different parts of the Chesapeake Bay Watershed and/or for different climate scenarios (or simultaneous changes in different CO_2 or weather factors)? In some ways, the paper might be seen as a case study. More explanation of why these two watersheds can allow us to draw broader conclusions beyond them could help to address this issue.

➔ The results of this study have implications for agricultural watersheds on coastal areas. Our analysis fully considers climate change impacts on croplands (crop growth, water and nutrient cycling) and their transport mechanisms (we referred this as “internal” watershed response) with considering detailed agricultural management practice. The two watersheds showed the typical site characteristics in the coastal plain, in terms of topographic and soil characteristics, and the agricultural practices considered in simulation are commonly used in the CBW region. Hence, the findings from this study can be applicable to other catchments in the CBW region. We will highlight this implication on section 4 in the revised manuscript.

2.3. Third, including the statistical analyses is a nice idea, but it is important to ensure that the tests are appropriate. Do these samples meet the assumptions of the tests that were used (such as independence)?

➔ We improved our statistical analysis to address the issue raised by a reviewer as below:

We conducted a statistical analysis to test if the changes in hydrologic variables by climate variability and change were significantly different from those under the baseline scenario. Note that we used monthly outputs (168 samples) for this analysis. We used both parametric and nonparametric methods to run our tests to avoid any problem caused by not meeting statistical assumptions for different tests. The statistical significance for the difference was indicated by *p-value*.

Specific comments:

2.4. Abstract: Perhaps mention the analysis of crop growth changes in the abstract?

➔ As suggested, we will briefly mention crop growth in the abstract as followings:

Using SWAT model simulations from 2001 to 2014, as a baseline scenario, the predicted hydrologic outputs (water and nitrate budgets) and crop growth were analyzed at multiple temporal scales.

Crop biomass increased by elevated CO_2 concentration while it decreased by precipitation and temperature increases.

2.5. Might be good to include some discussion of: How representative of historic climate was 2001-2014? Or, more specifically, the calibration years of 2001-2008? Was any cross validation done to assess the sensitivity of the selection of these groupings and time periods?

➔ We did not conduct any analyses to select the calibration period. Due to unavailability of observations before 2001, the calibration and validation periods were set from 2001. However, the calibration period (2001 - 2008) likely include representative wet, dry, and average climate conditions as recommended by the model guideline (Arnold et al., 2012). Compared to the distribution of past 30-year annual precipitation data (1981 - 2010), 8-year precipitation data over calibration period fully accounted for three representative climate conditions (Figure 1). However, validation period tends to include wet conditions.

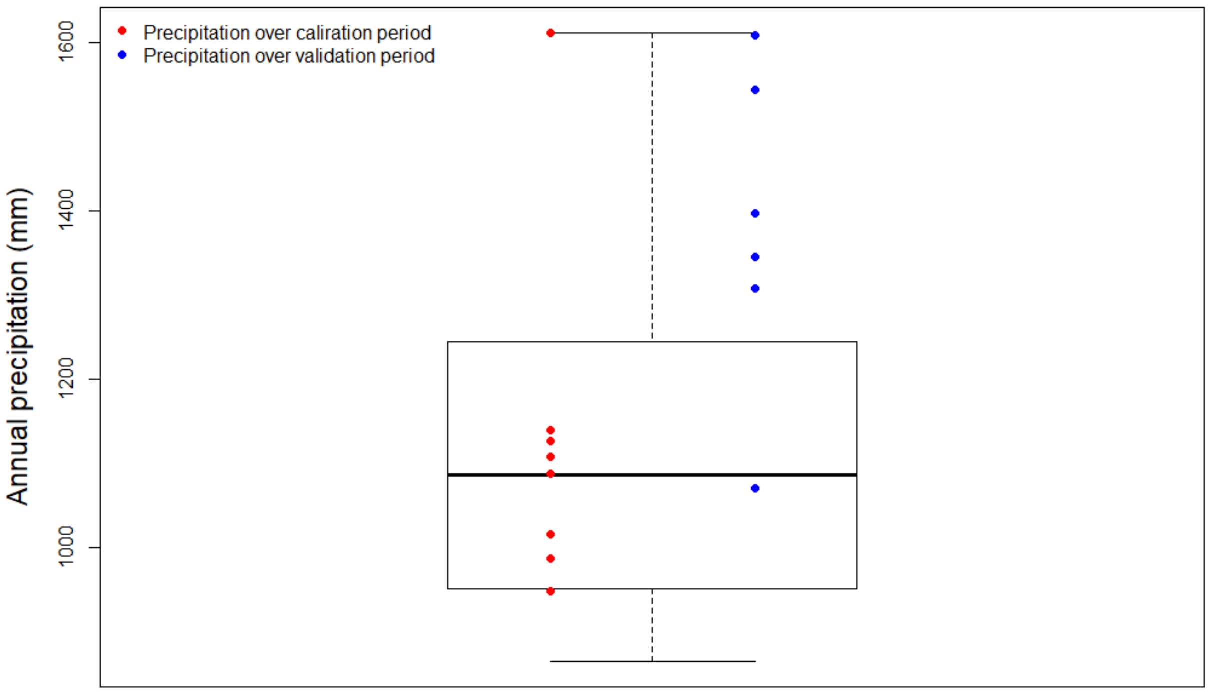

Figure 1. Comparison annual precipitation between past 30 years and calibration/validation. Note: The box plot was drawn using annual precipitation over past 30 years (1981 - 2010). The red and blue dots indicate annual precipitation over calibration and validation periods, respectively.

➔ We will state the brief discussion of climate conditions over the calibration and validation periods in the revised manuscript as below:

It should be noted that due to unavailability of observations before 2001, model calibration and validation were initiated from 2001. Compared to past 30-year precipitation data (1981 - 2010), climate condition over the calibration period (2001 - 2008) was shown to include representative wet, dry, and average climate conditions while the validation period (2009 - 2014) was dominant by wet conditions.

2.6. How were the two levels of increase in temperature and precipitation selected? From results in Najjar et al 2009?

➔ Based on the results from Najjar et al. (2009), two levels were selected. We improved the description of climate sensitivity scenarios in the revised manuscript as below:

We used the maximum increase rate (and value) for 2040 – 2069 (precipitation: 11 % and temperature: 2.9 °C) and 2070 – 2099 (precipitation: 21 % and temperature: 5.0 °C) to set the precipitation and temperature sensitivity scenarios. For example, the baseline precipitation increased by 11 % and 21 % for Scenario 3 and 4, respectively, and 2.9 °C and 5.0 °C were added to the baseline temperature for Scenario 5 and 6, respectively (Table 4).

2.7. Likely impact of using humidity, wind speed and solar radiation from the built in weather generator? Is this commonly done?

➔ When those three climate values were available from GCM data, a weather generator has been widely used in previous studies (Jayakrishnan et al., 2005; Ficklin et al., 2009; Wallace et al., 2017). Wang et al. (2009) also stated that *"the use of weather generators for downscaling monthly GCM data is actually not uncommon"*. Therefore, use of a weather generator can be regarded as one of potential ways to prepare humidity, wind speed, and solar radiation.

2.8. How much nitrate data was used and/or how often were the nitrate grab samples taken? Are there studies assessing the accuracy of using USGS LOAD ESTimator?

➔ The LOADEST is used commonly to generate continuous data from grab sample data (Lee et al., 2016). We used 133 samples to make continuous monthly data over the simulation periods of 168 months. Jha et al. (2013) reported that the LOADEST performed well in predicting water quality variables (e.g., nitrogen and phosphorus) with R^2 ranging from 0.97 to 99. This point will be addressed in the revised draft as below:

The USGS LOAD ESTimator (LOADEST, Runkel et al. (2004)) was used to generate continuous monthly nitrate loads from nitrate grab sample data (133 samples over the simulation period) that were obtained from the Chesapeake Bay Program (CBP, TUK#0181) for the TCW, and obtained from USGS gauge station data (USGS#01491000) for the GW. The LOADEST is used commonly to generate continuous data from discrete data and it was shown to accurately generate water quality variables (Jha et al., 2013; Lee et al., 2016b).

2.9. How was the 2-year warm-up period used in the SWAT modeling?

➔ The simulation started from 1999 to 2014 using observed precipitation and temperature (humidity, solar radiation, and wind speed were generated from a weather generator). The simulation results from 2001 to 2014 were only analyzed and the simulations over the 2-year warm-up period was not considered. The warm-up period is generally set to achieve equilibrium states and the model outputs are more reliable when setting up the warm-up period (Rahman and Lu, 2015).

2.10. Good that a number of statistics were used to assess model performance. Since NSE in real space more heavily weights the larger flow values, how well were the low flows captured? (Estimating NSE of the natural logarithms of the streamflows can also be helpful for this.)

➔ As suggested, we calculated the NSE for the natural logarithm of stream flow to evaluate the model predictability on low flows (Kiptala et al., 2014). Model performance measures indicate "Satisfactory" to "Very Good" for the two watersheds as shown in the table below. Therefore, low flows were also well depicted by our calibrated model.

Table 1. NSE for the natural logarithm of stream flow

Period	Variable	Stream flow	
		TCW	GW
Calibration	NSE	0.828***	0.719**

Validation	NSE	0.556*	0.727**

Model performances were rated based on the criteria of Moriasi et al. (2008); * Satisfactory, ** Good, and *** Very Good; Satisfactory (0.5 < NSE ≤ 0.65), ** Good (0.65 < NSE ≤ 0.75), and *** Very Good (0.75 < NSE ≤ 1.0).

➔ We will briefly discuss this analysis in the method and result sections as below:

Method section*: We also calculated NSE for the natural logarithm of stream flow to evaluate the model predictability for low-flows (Kiptala et al., 2014).*

Result section*: The model performance measures for low-flows (NSE for the natural logarithm of stream flow) also indicated "satisfactory" to "very good" (Table 5).*

➔ The NSE for the natural logarithm of stream flow will be added to Table 5.

2.11. Line 324-327: **1)** I'm a little unclear on this method and what ensemble is referring to here Are you taking the average across the whole time period predicted? Or are there multiple simulated outputs per monthly, seasonal, annual time period? "The range of changes in simulated outputs was represented with the ensemble mean to show overall responses of watershed hydrological processes to climate change (Shrestha et al., 2012; Van Liew et al., 2012)." **2)** Also with regards to the 95 PPUs estimated – some more explanation of the sample of simulations used would be helpful.

➔ **1)** With new GCM data, we calculated the ensemble mean by averaging the delta-change values of the five GCM projections with equal weight.

We will illustrate this process in the revised manuscript as below:

We calculated the ensemble mean by averaging the delta-change values of the five GCMs with equal weight because substantial variations existed among the GCM projections (Shrestha et al., 2012; Van Liew et al., 2012). Then, the SWAT model was simulated using the ensemble mean to predict hydrological processes under future climate conditions.

➔ **2)** We will improve the description of 95 PPU in the revised manuscript as below:

The 95 PPU was computed based on all simulated outputs generated during the calibration process (1,000 sets). The 95 PPU was represented as the range of values between the 2.5 and 97.5 percentiles of the cumulative distribution of simulated outputs.

2.12. Line 361-363: specify what "good" or "very good" meant numerically or list some numbers from the table.

➔ We will briefly specify the numerical meanings of model performances (e.g., Satisfactory, Good, and Very Good) only using NSE in the revised manuscript because the note of Table 5 fully illustrates the numerical meanings of each performance.

2.13. Figure 3: Do you know why there is such a difference between the two watersheds in terms of the 95 percent prediction uncertainty?

➔ This was likely due to the difference in soil characteristics between the two watersheds. The TCW and GW are dominated by well- and poorly-drained soils, respectively, and therefore “groundwater” is the major water transport pathway for the TCW while “surface runoff” is for GW.

Hence, our calibration shows TCW was more sensitive to the parameters pertaining to “groundwater flow” (ALPHA_BF, GW_DELAY, GW_REVAP, RCHRG_DP, and GWQMN; see Table 3) but GW was more sensitive to the parameters related to “surface runoff” (e.g., CN2 and SURLAG; see Table 3). As these parameters were calibrated in different allowable ranges, the uncertainty bands for two watersheds were naturally different.

2.14. Figure 4: perhaps connecting the ET with a line would help? It’s a bit difficult to interpret

➔ Yes. The dotted graph will be changed to a line graph as suggested.

2.15. Line 377: Since you are presenting p-values, do these predictions meet the assumptions of the statistical tests?

➔ Please see the answer 2.3 – Note that tests were done with sufficiently large sample using both parametric and nonparametric methods.

2.16. Figure 5: Wouldn’t CO2 and temperature likely both increase simultaneously? How would this effect plant growth?

➔ As answered in 2.1, this paper examined the individual impacts of CO_2, temperature, and precipitation investigate the individual effects of the key climate factors on the crop biomass, water and nitrate cycling. And this paper disregarded the combinations of two or three climate sensitivity scenarios because those combinations cannot provide foreseeable changes and complete climate change/variability information (including seasonal and inter-decadal variability). Therefore, analyzing simultaneous increases in CO_2 and temperature is the beyond the scope of this study.

2.17. Figure 6: Since these are relative to the baseline, consider plotting pluses and minus relative to that value to better illustrate the changes?

➔ One of goals in this paper is to compare water and nitrate transport patterns between two watersheds. Therefore, visualizing absolute values for each pathway can better represent the difference between the two watersheds in terms of the major pathway for water and nitrate fluxes.

2.18. Section 3.3: Did the GCM model runs include changes in CO2?

➔ Yes. We set CO_2 concentration of 936 ppm for the GCM data as stated in the section 2.5.2.

2.19. Line 492-497: Should this section be sooner as it also impacts the results presented previously for the one-by-one simulations?

➔ We will divide the paragraph into two and put the paragraph explaining the overestimation of CO2 impacts in the SWAT model in the section 3.2.1 as suggested. The other paragraph explaining its potential impacts on the GCM results will remain in the section 3.3.2.

Technical corrections:

2.20. Line 22-23: Should the first line of the abstract perhaps read "be exacerbated by" rather than "exacerbate under"?

➔ It will be changed in the revised manuscript as suggested.

2.21. Line 62-63: The Chesapeake Bay is the largest estuary in North America and thus the US, not just within the mid-Atlantic region. Maybe this sentence could be restructured along the Line of: "Located in the Mid-Atlantic region, the Chesapeake Bay (CB) is the largest and most productive estuary in the United States (US)."

➔ The sentence will be changed as suggested in the revised manuscript.

2.22. Line 112 - 115: These two sentences seem to be saying the same thing as one another (and reference the same papers) – maybe cut one of the sentences?

➔ The first sentence will be deleted in the revised manuscript.

2.23. Line 357: I would use the word "outside" or something similar rather than "beyond" which might imply higher than (when the reality is that predictions are lower).

➔ It will be changed to "outside" in the revised manuscript.

2.24. Line 521 Section 4: I think this should read "Implications" with an "s" at the end?

➔ It will be changed in the revised manuscript.

2.25. Line 568: typo: "five GCMs data"

➔ As requested by another reviewer, we revised the climate change scenario to GCM scenario and the word, "five GCM data", will be deleted in the revised manuscript.

<Cited references>

Ficklin, D.L., Luo, Y., Luedeling, E., and Zhang, M.: Climate change sensitivity assessment of a highly agricultural watershed using SWAT. J. Hydrol. 374 (1), 16-29, 2009.

Jha, B. and Jha, M.K.: Rating Curve Estimation of Surface Water Quality Data Using LOADEST. J Environ Prot. 4(8), 849-856, 2013.

Kiptala, J.K., Mul, M.L., Mohamed, Y.A. and Van der Zaag, P.: Modelling stream flow and quantifying blue water using a modified STREAM model for a heterogeneous, highly utilized and data-scarce river basin in Africa. Hydrol Earth Syst Sc. 18 (6), 2287-2303, 2014.

Lee, C.J., Hirsch, R.M., Schwarz, G.E., Holtschlag, D.J., Preston, S.D., Crawford, C.G. and Vecchia, A.V.: An evaluation of methods for estimating decadal stream loads. J Hydrol. 542, 185-203, 2016.

Mearns, L.O., Hulme, M., Carter, T.R., Leemans, R., Lal, M., Whetton, P., Hay, L., Jones, R.N., Kittel, T., Smith, J. and Wilby, R.: Climate scenario development, 2001.

Moriasi, D. N., Arnold, J. G., Van Liew, M. W., Bingner, R. L., Harmel, R. D., and Veith, T. L.: Model evaluation guideLine for systematic quantification of accuracy in watershed simulations. T. ASABE. 50(3), 885-900, 2007.

Najjar, R., Patterson, L., and Graham, S.: Climate simulations of major estuarine watersheds in the Mid-Atlantic region of the US. Climatic Change, 95 (1-2), 139-168, 2009.

Rahman, M.M. and Lu, M.: Model spin-up behavior for wet and dry basins: a case study using the Xinanjiang model. Water, 7(8), 4256-4273, 2015.

Shrestha, R.R., Dibike, Y.B. and Prowse, T.D.: Modelling of climate-induced hydrologic changes in the Lake Winnipeg watershed. J. Great Lakes Res. 38, 83-94, 2012.

Van Liew, M.W., Feng, S. and Pathak, T.B.: Climate change impacts on streamflow, water quality, and best management practices for the shell and logan creek watersheds in Nebraska, USA. Int. J. Agric. Biol. Eng. 5(1), 13-34, 2012.

Wallace, C.W., Flanagan, D.C. and Engel, B.A.: Quantifying the Effects of Future Climate Conditions on Runoff, Sediment, and Chemical Losses at Different Watershed Sizes. Trans. ASABE, 60(3), 915-929, 2017.

Wang, R., Kalin, L., Kuang, W. and Tian, H.: Individual and combined effects of land use/cover and climate change on Wolf Bay watershed streamflow in southern Alabama. Hydrol. Process. 28(22), 5530-5546, 2014.

Wolock, D.M. and McCabe, G.J.: Estimates of runoff using water-balance and atmospheric general circulation models. J. Am. Water Resour. As., 35(6), 1341-1350, 1999.