A predictive model for Lake Chad total surface water area using remotely sensed and modeled hydrological and meteorological parameters and multivariate regression analysis

Frederick Policelli [a,d], Alfred Hubbard [b,c], Hahn Chul Jung [a,c], Ben Zaitchik [d], Charles Ichoku [e]

[a] NASA Goddard Space Flight Center
8800 Greenbelt Rd
Code 617, Hydrological Sciences Lab
Greenbelt, MD, 20771, USA

[b] NASA Goddard Space Flight Center
8800 Greenbelt Rd
Code 618, Biospheric Sciences Lab
Greenbelt, MD, 20771, USA

[c] Science Systems and Applications, Inc.
10210 Greenbelt Rd
Lanham, MD, 20706, USA

[d] Johns Hopkins University
Dept. of Earth and Planetary Sciences

3400 N. Charles St.
Baltimore, MD 21218, USA

[e] Howard University
Dept. of Interdisciplinary Studies
1840 7th Street, NW
Washington, DC 20001, USA

* Correspondence: frederick.s.policelli@nasa.gov; Tel.: +01-301-614-6573

Abstract

Lake Chad is an endorheic lake in west-central Africa at the southern edge of the Sahara Desert. The lake, which is well known for its dramatic decrease in surface area during the 1970s and 1980s, experiences an annual flood resulting in a maximum total surface water area generally during February or March, though sometimes earlier or later. People along the shores of Lake Chad make their living fishing, farming, and raising livestock and have a vested interest in knowing when and how extensive the annual flooding will be, particularly those practicing recession farming in which the fertile ground of previously flooded area is used for planting crops. In this study, the authors investigate the relationship between lake and basin parameters, including rainfall, basin evapotranspiration, lake evapotranspiration, lake elevation, total surface water area, and the previous year's total surface water area, and develop equations for each dry season month (except November) linking total surface water area to the other parameters. The resulting equations allow the user to estimate the December average monthly total surface water area of the lake in late November, and to make the estimates for January to May in early December. Based on the results of a Leave One Out Cross Validation analysis, the equations for lake area are estimated to have an average absolute error ranging from 5.3 percent (for February estimates) to 7.6 percent (for May estimates).

[1]Keywords: surface water area, Lake Chad, precipitation, evapotranspiration, lake height

1. Introduction

Lake Chad is a shallow, endorheic lake in the Sahel region of west-central Africa shared by Chad, Nigeria, Niger, and Cameroon. There have been numerous hydrological models developed for Lake Chad. Each of these has limitations relative to the statistical model for lake area we present here.

Bader et al. (2011) developed a hydrological model that simulates the water level in the northern pool, southern pool, and archipelago using riverine and direct rainfall inputs to the lake. It also estimates total water area for each of the pools. According to (Lemoalle et al., 2012), the model results correspond well with satellite measurements of northern pool surface water area and with satellite measurements of the total surface water area from 1980 onward, though these

---

Abbreviations: Inter-Tropical Convergence Zone (ITCZ), above sea level (ASL), Climate Hazards Group InfraRed Precipitation with Station data (CHIRPS), Famine Early Warning Systems Network (FEWS NET) Land Data Assimilation System (FLDAS), Land Surface Temperature (LST), evapotranspiration (ET), Leave One Out Cross Validation (LOOCV), precipitation (P). For reading regression results: Pws is total south basin wet season precipitation, ETws is total south basin wet season evapotranspiration, LakeETws is wet season lake ET percent variation from the 1988 to 2016 wet season mean, H is November lake height variation from the 1993-2002 mean, A is the average surface water area for the given month, A- is the previous year's average surface water area for the given month.

results are not quantified. One disadvantage of this model is that it requires input of the Chari River and Komodougu-Yobe River discharges, data that is not publicly available.

Coe and Foley (2001) report the results of a hydrological model of Lake Chad. They describe a model with “good agreement with the inferred lake area” during simulations for the years 1954 to 1967. The “inferred lake area” is not a direct measurement of lake area, but rather is derived from a relationship between lake area and lake level. It is important to note that the hydrology of Lake Chad changed significantly in the 1970’s during its transition from “Normal Lake Chad” to “Small Lake Chad” and the model, based on a coarse (10 km resolution) digital elevation model, may not be adequate to define the lake after this transition.

Gao et al. (2011) developed a hydrological model of Lake Chad. They compared images of the lake extent from the model with images derived from remote sensing. Three image pairs were shown, for October 31, 1963; December 25, 1972; and January 31, 1987. Each pair of images from the two earlier dates (before the transition from “Normal Lake Chad” to “Small Lake Chad”) looked quite similar, though no numerical value was provided. The model-observation pair for the post transition period did not look very similar. This raises the question of the utility of the model for producing lake area in the current, post transition period.

Lemoalle (2004) developed a crude expression for Lake Chad surface area based on a simplified water balance model and described it as a “first approximation.” That model assumes no seepage from the lake and requires knowledge of the streamflow to the lake, which as previously noted is not publicly available. Delclaux et al. (2008) developed a hydrological model of the Lake Chad Basin, however the results they presented included streamflow and elevation, but not surface area.

Coe and Birkett (2004) used upstream measurement of river height along with in-situ stream flow and gauge height to estimate river discharge 500 km downstream and wet season height of Lake Chad, greater than 600 km downstream. Their method, though, clearly relies on hard-to-obtain in-situ measurements and does not include lake area.

The first objective of this paper is to assemble and examine a set of satellite- and model-based data for the southern Lake Chad Basin relevant to developing a statistical model for the area of the lake during its flooding season. This data set includes time series of satellite- and gauge-based precipitation and modeled ET for the southern part of the basin, modeled lake ET data, satellite-based lake elevation data, and satellite-based estimated lake total surface water area. Given the limitations of the existing models described above, the second objective is to develop a predictive statistical model for total lake surface water area using regression methods on the data set. The regression method includes backward elimination variable selection and a Leave One Out Cross Validation (LOOCV) analysis to optimize the resulting statistical model.

2 Study Area

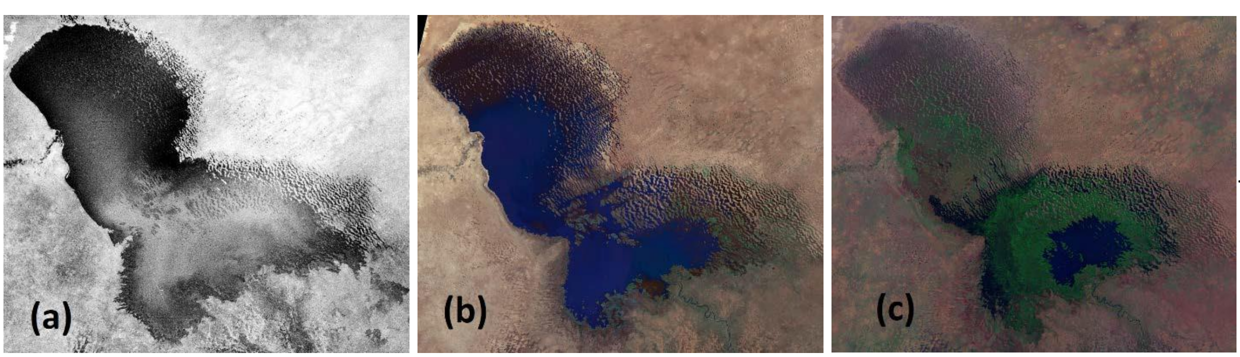
The Lake Chad basin is approximately 2.5 million square kilometers, about eight percent of the African continent, and the largest endorheic basin in the world (Gao et al., 2011). Lake Chad is the terminal lake of this basin. The northern part of the basin lies within the Sahara and does not generate runoff that reaches Lake Chad (Delclaux et al., 2008).

For this reason, we work with the southern part of the basin (figure 1).

Fig. 1 Southern Lake Chad Basin (white), lake, and major rivers (blue)

The lake's average depth varies between 1.5 and 5 m. Any change in lake volume translates to a substantial change in lake shoreline and area ("Lake Chad flooded savanna", World Wildlife Fund, no date, https://www.worldwildlife.org/ecoregions/at0904).

In the 1960's the lake's area was on the order of 25,000 sq. km.; in the mid-1980s its area was reported to be about one tenth of that size (Grove, 1996), though it is not clear if that includes flooded vegetation. If one includes flooded vegetation, the lake's annual peak area for 2017 is estimated at close to 14,700 sq. km (Policelli et al., 2018). Figure 2 shows the evolution of Lake Chad from the time of the earliest space-based images of the lake.



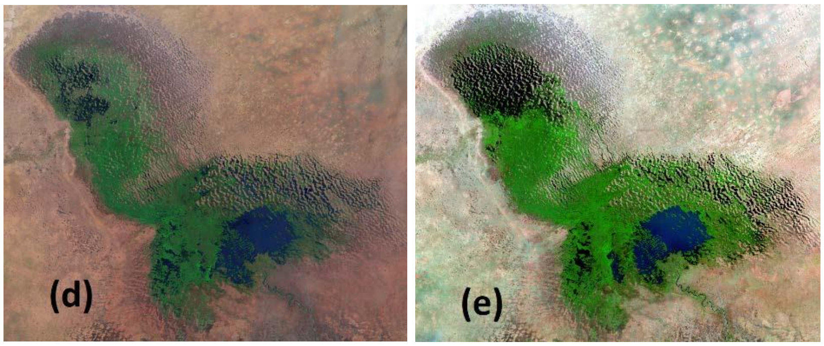


Fig. 2 Evolution of Lake Chad. Optical imagery from (a) 1963, Argon satellite (b) 1973, Landsat 1 (c) 1987, Landsat 5 (d) 2003, Landsat 7 (e) 2013, Landsat 8. Images provided by U.S. Geological Survey, Department of the Interior/USGS

Below about 280 m ASL, the lake separates into a southern pool and a northern pool divided by “the Great Barrier”, and the southern pool separates from the eastern “archipelago” of sandy islands (Lemoalle, 2004).

The population of the lake shore is around 2 million (Magrin, 2016) and the people make a living through a combination of fishing, farming, and raising livestock (Sarch and Birkett, 2000). “Recession farming” is an important method of farming in the region whereby farmers plant in the enriched soils following each year’s flood pulse. Because of the complexity of the hydrology, it is difficult to provide farmers with information on the timing of the floods and a sense of how large the flood is going to be in any given year. This can be a serious problem for farmers who grow crops near the lake shore and periodically lose crops to flooding (Okpara et al., 2016).

According to (Sarch and Birkett, 2000), the "farming start date" at villages on the south-west shore of Lake Chad begins in mid-January to late February. It seems reasonable to conclude that predictions of lake surface area made in late November or early December could be used for agricultural decision making for locations with similar start dates.

There are a number of dams and irrigation schemes in the river basins that drain into Lake Chad, such as those of the Chari-Logone and Komadougu-Yobe river systems. However, at the large scale, they do not amount to a substantial revision of the natural seasonal hydrological patterns, which are largely determined by the West African Monsoon and the position of the ITCZ (Birkett, 2000).

The surface of Lake Chad is not flat and level; the barriers to flow from the southern pool to the northern pool and archipelago result in the water in these areas frequently being at different elevations. Additionally, the local winds and flow of water from the Chari River into the lake (addressed further in the Discussion section) combined with the complex shape of the lake lead to an evolving lake surface (Carmouze et al., 1983). Lake surface elevation time series data used in this study refer to satellite-based measurements made in the relatively small open water portion of the lake in the southern pool.

Figure 3 presents areas of the lake at or below selected elevations. The topographic data used to create figure 3 is from a 1 arc-second (~30m) resolution Digital Elevation Model (DEM) from the NASA Shuttle Radar Topography Mission (SRTM). The blue areas would correspond to lake levels (limited by the accuracy of the DEM) if the lake surface were flat and the lake filled uniformly. However, since this is not the case, the areas should only be viewed as providing context for Lake Chad landforms and hydrology.

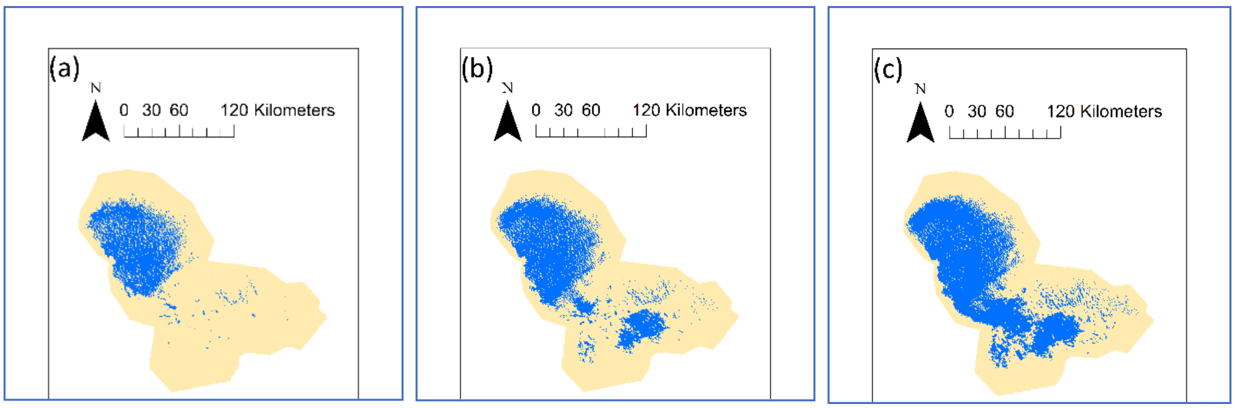


Fig. 3 Blue areas have elevations at or below (a) 279m ASL, (b) 280m ASL, (c) 281m ASL

According to (Leblanc et al., 2011) most of the flooded area of Lake Chad is covered by aquatic vegetation including rooted and floating plants. This area is not readily measured with optical remote sensing, but must be accounted for to get an accurate estimate of total surface water area for the lake (Policelli et al., 2018).

Lake Chad receives 90-96% of its water from the Chari-Logone River system (Zhu et al., 2017a), with the remaining coming from smaller tributaries and direct rainfall. Most of this water arrives from such a distance that the peak lake level and lake area occur months after the rainfalls that produce them. The delay is due to the slow runoff and routing of flood water to Lake Chad from the southern portion of the basin where precipitation rates are highest (Leblanc et al., 2011). It is the reason that the Lake Chad peak level and area take place in the dry season.

3 Data and Methods

3.1 Data

The key datasets used for this research are (1) Climate Hazards Group Infrared Precipitation with Station data (CHIRPS), (2) satellite altimetry-based lake surface elevation data, (3) evapotranspiration (ET) data from the Famine Early Warning Systems Network (FEWS NET) Land Data Assimilation System (FLDAS), (4) percent of 1988-2016 lake ET average estimates, (5) the HydroBASINS shape file for the southern Lake Chad Basin extent and (6) lake total surface water area data, described in (Policelli et al., 2018).

3.1.1 CHIRPS (Funk et al., 2015) is a quasi-global precipitation dataset produced by the Climate Hazards Group at the University of California, Santa Barbara using both satellite data and rain gauge data. We used the high resolution (.05 degree x .05 degree) daily Africa rainfall dataset, which is available from January, 1981 and updated through the previous month around the third week of each month. In comparison with the lower resolution gauge-based GPCC reference precipitation product (Schneider et al., 2015), for wet seasons in Africa CHIRPS has a mean error of 79 mm per 3 months, a mean bias of 0.22 and a correlation of 0.56 (Funk et al., 2015). In comparison with other observations-based precipitation products for Africa, CHIRPS data has higher spatial resolution, better coverage of rain gauge stations, and applies improved statistical methods (Badr et al., 2016).

3.1.2 Lake surface elevation data are provided by the Global Reservoir and Lake Monitor (G-REALM) on the USDA Crop Explorer website ("Satellite Radar Altimetry: Global Reservoir and Lake Elevation Database," no date, https://www.pecad.fas.usda.gov/cropexplorer/global_reservoir/). Lake surface elevation data are produced from the radar altimetry satellite missions Topex/Poseidon (1992-2002), Jason-1 (2002-2008), Jason-2 (2008 to 2016), and Jason-3 (2016 to present). The altimeters have a "footprint" diameter ranging from about 200m to several kilometers depending on the target's surface roughness. Each of the altimeters used for Lake Chad elevation data have a ten-day repeat time. The accuracy of the lake surface elevation data for Lake Chad is approximately 0.29 m (Ricko et al., 2012). Lake surface elevation data are provided as the variation from the 1993-2002 mean height.

3.1.3 The FLDAS ET data product is based on the Noah 3.3 Land Surface Model's total ET, which is the sum of bare soil evaporation, evaporation of water intercepted by the canopy, and transpiration, weighted by the coverage fraction of each component (McNally et al., 2017). The spatial resolution of the FLDAS ET is 0.1 degree and we are using monthly data. The FLDAS ET data is available from October 1992 to the present. FLDAS ET has been evaluated against estimates from the Simplified Surface Energy Balance (SSEBop) satellite data-based model (Senay et al., 2013). This evaluation indicates that FLDAS and SSEBop ET have significant but limited correlation ($r < 0.5$) for percentage ET variations in West Africa. According to McNally et al., 2017, it is not entirely clear why the correlation between the FLDAS-ET and SSEBop-ET is somewhat poor in West Africa, though it may have something to do with instability of the SSEBop algorithm in that area.

3.1.4 We estimated the Lake Chad wet season ET as a percentage of the 1988-2016 wet season average. Because most (approximately constant at nearly 90 percent by our estimation) of the lake consists of flooded vegetation, and multiple types of vegetation are present, it is a very difficult research problem to estimate the total monthly ET from Lake Chad. However, because the ratio of open water to flooded vegetation is roughly constant over time, we were able to estimate the lake ET percent of average as the ratio of the open water evaporation for the full extent of the lake to the average open water evaporation for the full extent of the lake for 1988-2016. We used the Complementary Relationship Lake Evaporation (CRLE) model (Morton, 1986) for estimates of lake open water evaporation. The meteorological forcing data for the CRLE model were provided by NOAA NCEP-DOE Reanalysis 2. The main validation work by the developers of the CRLE model was to compare model results with lake evaporation from water balance analyses for seventeen selected lakes. On an annual basis, the model results were within a maximum of seven percent of the water balance results. Monthly results suffered an unspecified degradation of accuracy (Morton, 1986).

3.1.5 The World Wildlife Fund (WWF) HydroBASINS (Lehner, 2013) provided the shape file for the southern Lake Chad Basin, which was used to calculate rainfall and ET for the portion of the basin that generates very nearly all of the runoff that reaches the lake. The HydroBASINS global database of basin shapes was developed using the WWF HydroSHEDS data (Lehner et al., 2011), which has approximately 500 m resolution. The HydroSHEDS product was developed using SRTM data. No validation description or accuracy assessment was found for HydroBASINS.

3.1.6 Lake Chad total surface water area data from the research for (Policelli et al., 2018) were derived from (1) 1 km resolution Land Surface Temperature (LST) data from the NASA Terra satellite's MODIS sensor and adjusted using ESA C-band radar data from Sentinel-1a, and (2) total Lake Chad surface water area estimates by (Leblanc et al., 2011) derived from 5 km resolution LST data from the Meteosat MVIRI sensor and bias corrected by (Policelli et al., 2018). The LST method for estimating lake surface water area is based on the fact that the lake, including flooded vegetation, is cooler than the surrounding landscape during the day (Leblanc et al., 2011). Monthly average area was used for this research. This data was produced for the 1988-1989 dry season through the 2016-2017 dry season. MODIS LST data with cloud cover greater than five percent were not used in the development of the lake area data. The validation done for this data was comparison of the two datasets used to create the lake area time series. The estimated lake areas for the two products during the period of overlap were within approximately three percent of each other.

3.2 Methods

Monthly precipitation and ET were calculated for the southern portion of the basin from October 1992 to May 2017 using the HydroBASINS shape file for this area. The Lake Chad total surface water area time series was then checked for correlation with P–ET using a series of time lags to find the peak correlation. It was expected that net precipitation (P-ET) would be more closely correlated to lake area than either of the components. Similarly, the total lake surface water area was checked for correlation with the lake elevation data, and it was expected that a close correlation would exist between these variables (Busker et al., 2018). Additionally, the total lake surface water area was checked for correlation with the lake elevation data, the precipitation data and the ET data for the southern portion of the basin, the lake ET percentage of 1988-2016 average lake ET, and the previous year's total lake surface water area.

Next, a multivariate regression analysis was performed using the datasets to establish equations linking total surface water area for a given month to one or more of the other (independent) variables. In order to find the best relationship between the lake area and the independent variables, we used several methods of regression: 1st order linear regression, 2nd and 3rd order polynomial regression, and the linear-log method of regression. We also used the backward elimination method of variable selection to optimize the equations. To mitigate the risk of overfitting the data, we used the "rule of thumb" of ten data points for each variable included in the final equation (Harrell et al., 1996), except for 1st order linear regression, in which case as few as two data points for each variable included is permitted (Austin and Steyerberg, 2015). However, we did not use less than six data points for each variable in our 1st order linear regression analysis.

The current year's total surface water area for each month during the dry season was the dependent variable. The independent variables were (1) total wet season precipitation for the southern Lake Chad Basin, (2) total wet season ET for the southern Lake Chad Basin, (3) wet season lake ET as a percentage of the 1988-2016 average lake ET, (4) November (typically the highest) variation from the 1993-2002 mean lake surface height, and (5) the total surface water area of the given month for the previous year. A Leave One Out Cross Validation (LOOCV) was performed for the regression analysis, in which one data point was left out, the equation was generated using regression, and the equation was evaluated for the one left out data point. This was repeated for each yearly data point and an average absolute value of the percent error was determined for the total dataset. For comparison, the average absolute value of the percent error was also determined using the average total surface water area for a given month as the prediction, and using the previous year's area as the prediction for a given month.

The study covered the time period from October 1992 to May 2017 which is the intersection of the period of available FLDAS evapotranspiration data and the available total lake surface water area data. Three years (2006-2008) were excluded from the study because of insufficient lake surface elevation data.

4. Results

4.1 Datasets

Total monthly precipitation and ET for the southern portion of the Lake Chad Basin were calculated using CHIRPS and FLDAS, respectively. Figure 4.a. shows the total monthly precipitation and ET for October 1992 through May 2017. The trend lines indicate essentially no change over this time period. Figure 4.b. shows the average monthly distribution of the precipitation and ET for the southern portion of the basin for October 1992 through May 2017. June through October is usually considered the wet season and November through May is considered the dry season (figure 4.b and Leblanc et al., 2011).

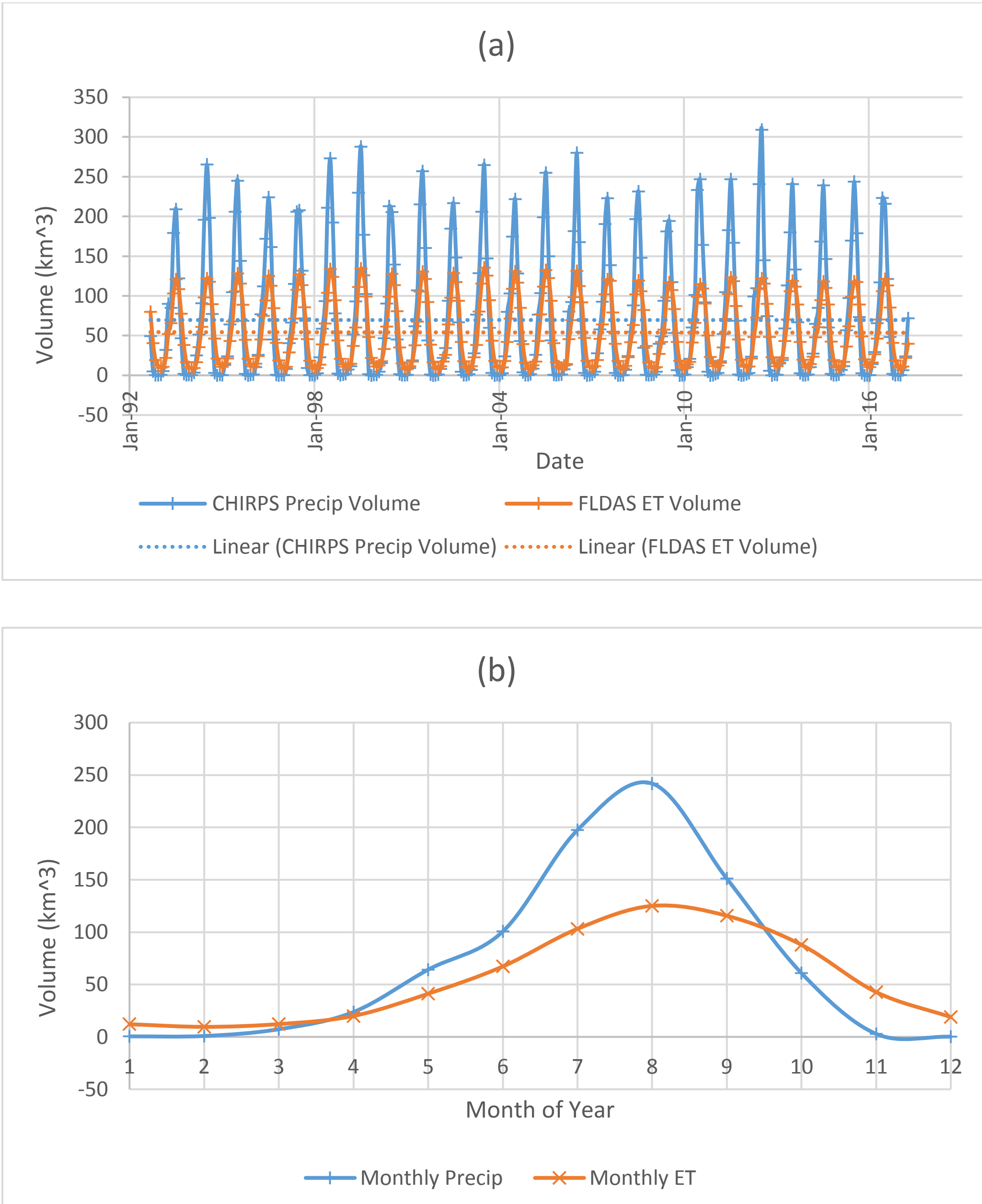


Fig. 4 a. monthly precipitation and ET in the Southern portion of the Lake Chad Basin b. average monthly precipitation and ET in the Southern portion of the Lake Chad Basin

Figure 5 shows the percent of Lake Chad wet season average ET from 1988 to 2016. The data exhibits an oscillatory behavior with a multi-year cycle and the trend is essentially flat.

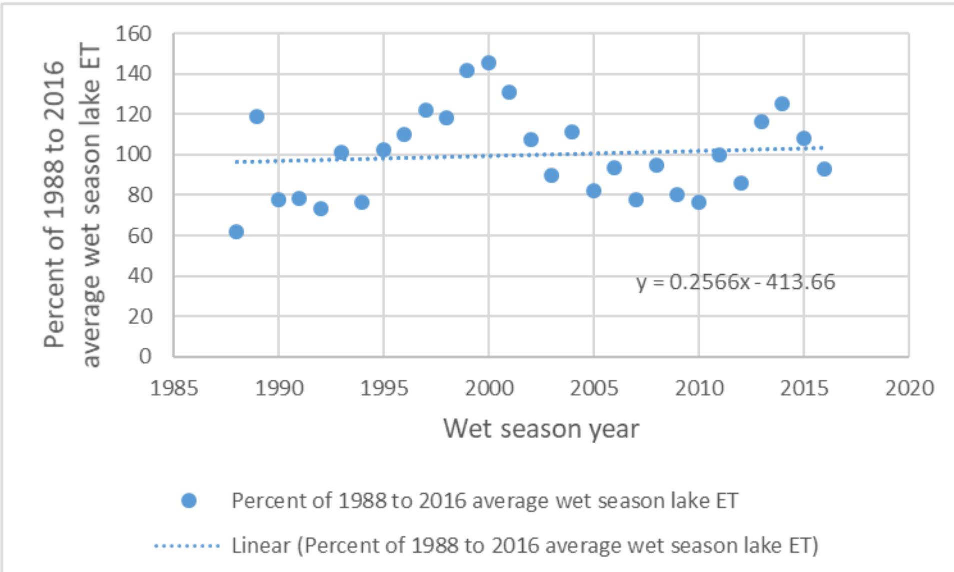


Fig. 5 Percentage of 1988 to 2016 average wet season lake ET

Figure 6.a. shows the lake elevation anomaly relative to the 1993-2002 mean level for October 1992 to May 2017. The trend line for July (the only month with complete data; not shown) shows a slight level of decrease with high variability. Figure 6.b. shows the average monthly lake elevation anomalies with respect to the 1993-2002 mean level for October 1992 through May 2017. Note that the average level is highest in November (during the dry season) and lowest in June (during the wet season).

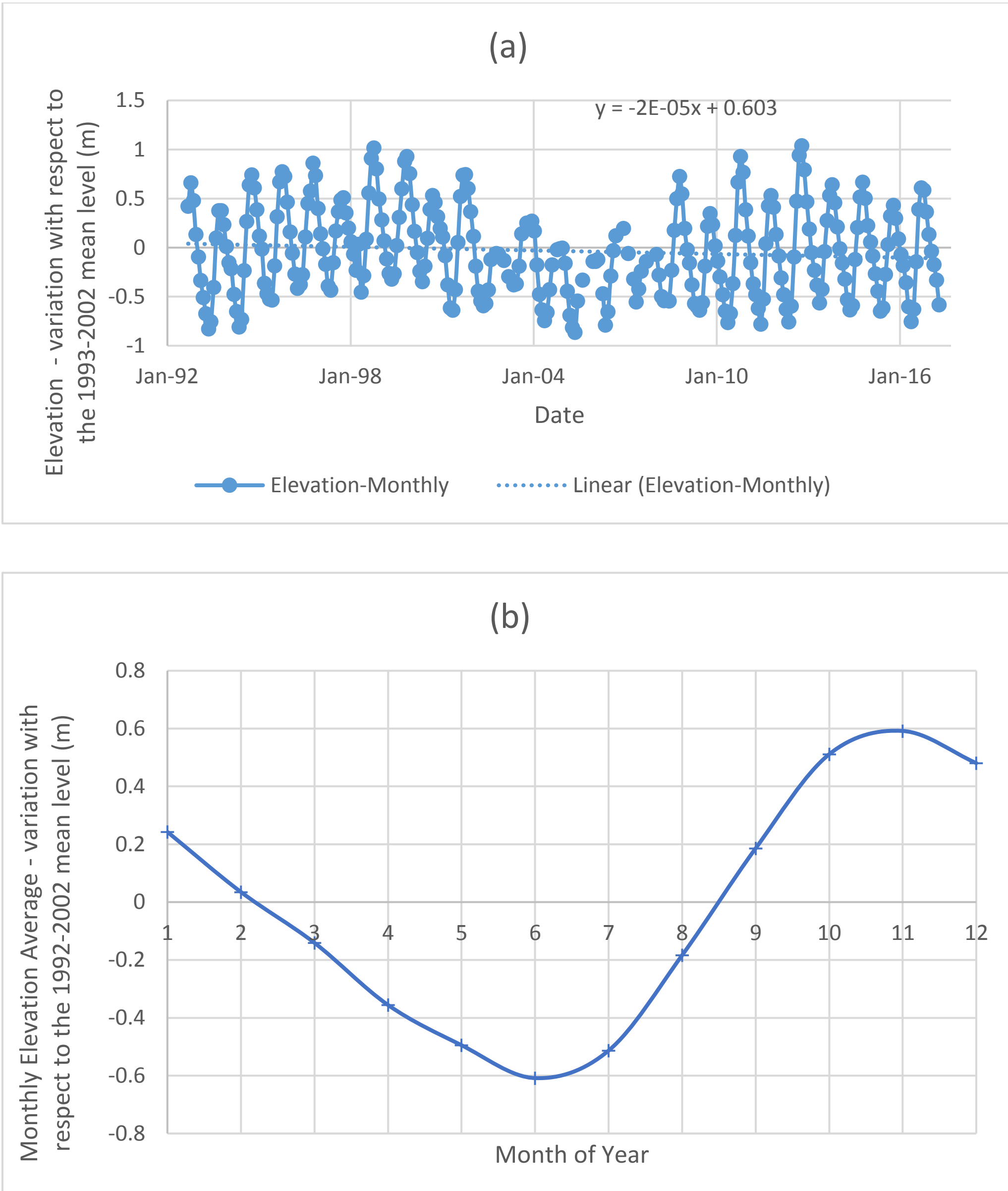
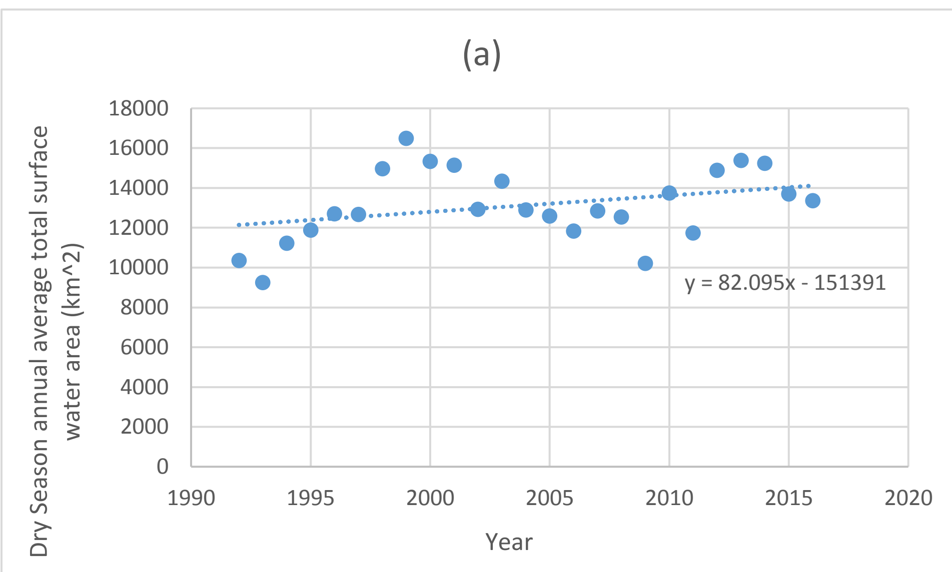


Fig. 6 a. lake elevation anomaly relative to the 1993-2002 mean level

b. average monthly lake elevation anomaly relative to the 1993-2002 mean level

Figure 7.a. presents the annual average dry season surface water area for Lake Chad from

November 1992 through May 2017. As expected, the pattern of the graph is similar to that of the graph of Lake Chad percentage of 1988 to 2016 average wet season lake ET (figure 5). The trend line shows an average increase of approximately 82 square kilometers per year in lake area over this period. Wet season data was not considered suitable for calculation of water extent because of cloud coverage and the fact that the LST method cannot distinguish well between soil moisture and flooded areas (Policelli et al., 2018). The total surface water area includes both open water and flooded vegetation. Figure 7.b. presents the average dry season monthly total surface water area for Lake Chad from November 1992 through May 2017. While the peak water elevation occurs on average in November, the peak total surface water area occurs in March on average. The peak total surface water area is not a sharp peak; the two antecedent and two following months are not far from the maximum



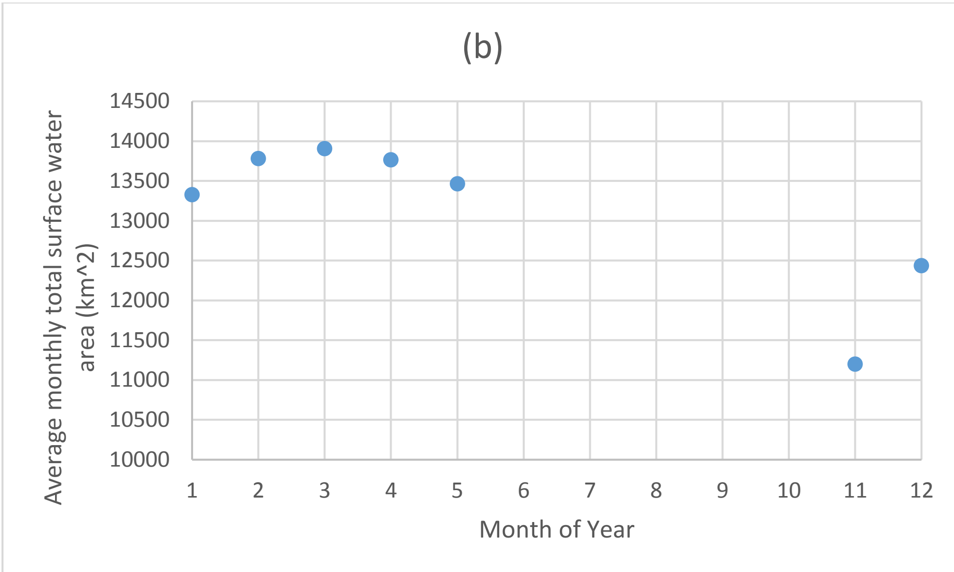


Fig. 7 a. average annual dry season water extent for Lake Chad b. average monthly lake water area for the dry season (November 1992 through May 2017)

4.2 A predictive model for Lake Chad total surface water area

During this part of the research, we asked the question: "for a given month, how well can we predict the total surface water area of Lake Chad"? To address this question, we investigated regression of the data to generate equations linking the independent variables to the dependent variable. Specifically, we looked at 1st order linear equations (Equation 1), 2nd order polynomial equations (Equation 2), 3rd order polynomial equations (Equation 3), and linear-log equations (Equation 4).

Equation 1: $A = a + b \cdot ETws + c \cdot LakeETws + d \cdot Pws + e \cdot H + f \cdot A$-

Equation 2: $A = a + b \cdot ETws + c \cdot ETws^2 + d \cdot LakeETws + e \cdot LakeETws^2 + f \cdot Pws +$

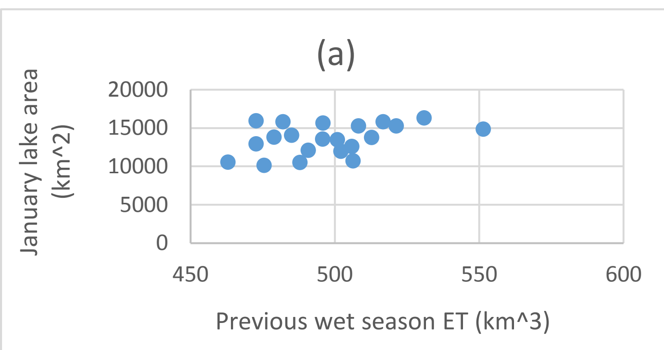
$$g \cdot Pws^2 + h \cdot H \; + j \cdot H^2 + k \cdot A\text{-} + m \cdot A\text{-}^2$$

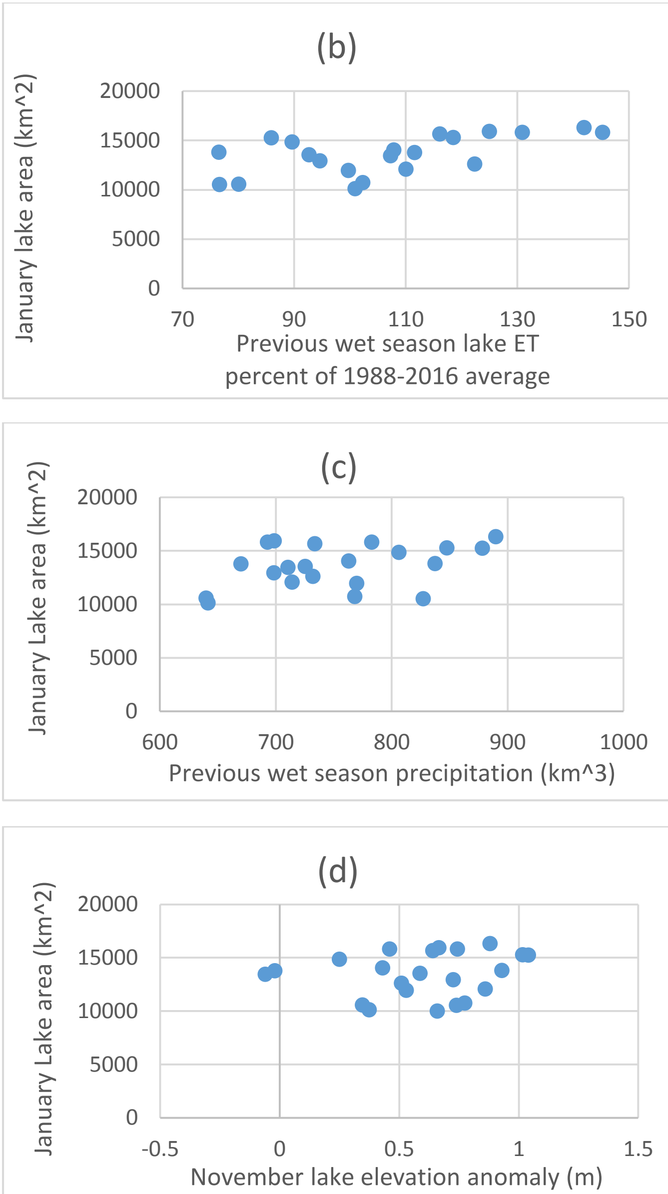
Equation 3: $$A = a + b \cdot ETws + c \cdot ETws^2 + d \cdot ETws^3 + e \cdot LakeETws + f \cdot LakeETws^2 + g \cdot LakeETws^3 + h \cdot Pws + j \cdot Pws^2 + k \cdot Pws^3 + m \cdot H + n \cdot H^2 + p \cdot H^3 + q \cdot A\text{-} + r \cdot A\text{-}^2 + s \cdot A\text{-}^3$$

Equation 4: $$A = a + b \cdot \log(ETws) + c \cdot \log(LakeETws) + d \cdot \log(Pws) + e \cdot \log(H) + f \cdot \log(A\text{-})$$

where: $A = A_{\text{Lake Chad, t}}$, $ETws = \sum_{June}^{Oct.}$ basin ET, $LakeETws = \sum_{June}^{Oct.} lake$ E/ lake $E_{1988\text{-}2016\ \text{avg}}$, $Pws = \sum_{June}^{Oct.}$ basin P , H = Lake $\text{Elevation}_{\text{Nov}}$ – Lake $\text{Elevation}_{1993\text{-}2002\ \text{avg.}}$, and $A\text{-} = A_{\text{Lake Chad, t - 1 year}}$

Figure 8 shows the scatter plots for each of the independent variables versus the average lake area for January.




(b)
January lake area (km^2)
20000
15000
10000
5000
0
70
90
110
130
150
Previous wet season lake ET
percent of 1988-2016 average
(c)
January Lake area (km^2)
20000
15000
10000
5000
0
600
700
800
900
1000
Previous wet season precipitation (km^3)
(d)
January Lake area (km^2)
20000
15000
10000
5000
0
-0.5
0
0.5
1
1.5
November lake elevation anomaly (m)

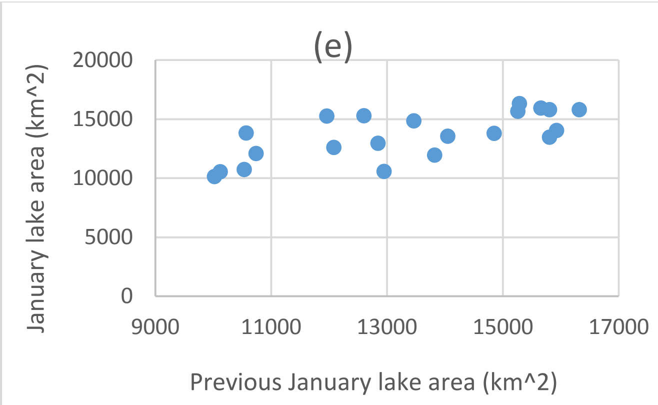


Fig. 8 Example (January) scatter plots for independent variables vs. lake area (a) previous wet season ET, (b) previous wet season lake ET percentage of the 1988-2016 average, (c) previous wet season precipitation, (d) November lake elevation anomaly, (e) previous January lake area.

From our regression analysis we found that for the equation types we examined, 1st order linear equations provided the minimum average absolute percent error from LOOCV for December, February and May. For January, March and April, higher order polynomial equations provided slightly lower (between 0.7% and 1.1% lower) LOOCV average absolute percent errors. However, because of the lack of a physical explanation for some of the higher order terms (A^{-2} and A^{-3} for instance), and the marginal gain for using the higher order equations, we decided to maintain consistency across months and use optimized 1st order linear equations for all months. Linear-log solutions for the regression analysis provided higher LOOCV average absolute percent errors than the linear 1st order and polynomial solutions. The final equations are provided in Table 1.

Table 1 Equations and performance metrics for Lake Chad total surface water estimated area in terms of ETws, LakeETws, Pws, H, A-

Forecast Mo.	Lake Area Equation	LOOCV error (%)	LOOCV std (%)	Adjusted R-Squared
December	-7064.38 + 23.53 * LakeETws + 12.57 * Pws + 0.62*A-	6.05	4.60	0.78
January	-11641.42 + 25.59 * ETws + 3364.12 * H + 0.78 * A-	6.90	4.91	0.76
February	-8443.93 + 14.40 * Pws +1771.05 * H +0.76 * A-	5.25	4.66	0.80
March	-7042.42 + 14.36 * Pws + 2074.06 * H + 0.65 * A-	7.20	5.95	0.69
April	-5187.30 +12.75 * Pws + 2249.73 * H + 0.59 * A-	7.53	5.14	0.64
May	-7402.54 + 17.15 * Pws + 2203.80 * H + 0.50 * A-	7.61	6.03	0.70

The result is that this model, using three variables for each month, can be used in late November (when the precipitation data is available) to predict the Lake Chad total surface water area for December, and in early December, (when lake elevation data is available for November) the model can be used to predict the total surface water area for January through May, with the expectation of between 5.3 and 7.6 percent error on average. This compares with (Table 2) the set of average absolute percent errors if the average value of the total water surface area for each month is used as a prediction,

Table 2 Average absolute percent error with average value used as prediction.

	average absolute % error
December	14.1
January	13.8
February	13.0

March	12.4
April	11.8
May	13.1

and (Table 3) the set of average absolute percent errors if the previous year's total water surface area for the given month is used as a prediction.

Table 3 Average absolute percent error with previous year's lake area used as prediction

	average absolute % error
December	10.0
January	9.0
February	10.6
March	12.5
April	12.8
May	15.6

For each of the months December through May, the regression approach provides a lower average absolute percent error than either the average value used as a prediction, or the previous year's area used as a prediction. As an example, for May, the LOOCV average absolute error is

7.6%, while the error using the average May area is 13.1% on average, and the error when using the previous year's area for May is 15.6% on average. Using the average May lake area of 13,363 sq. km, these percentages are equivalent to 1017 sq. km, 1751 sq. km, and 2085 sq. km respectively.

5. Discussion

The lake elevation data show that the maximum average monthly elevation typically occurs in November (though on occasion in October, and once during our record in January). It is curious therefore that the maximum area typically though not always occurs in February or March. The reason for this seems to be the complex meteorology and hydrology of Lake Chad. The movement of the water in Lake Chad is influenced by both the winds and the Chari-Logone water supply. Monsoon winds drive the displacement of the southern waters toward the north, and movement begins around the northeastern end of the Great Barrier in June, at the end of the low water. The Chari-Logone flood waters begin in August and provide half of their water in October and November when the northeasterly wind known as the Harmattan drives the water back toward the southern pool. This is also when the satellite radar altimeters (which collect data over the southern pool) typically record the highest levels. During the peak of the riverine flooding, water again reaches the northern pool and also spreads into the archipelago from the south basin. Following the end of the movement of water to the north pool in January, there is a general spreading of water in the southern pool to the periphery until April. (Carmouze et al., 1983). These movements result in a complex and changing lake surface topography, and are likely the reason we find poor correlation between lake elevation and lake area.

A full parameterization of the Lake Chad system, as implemented by (Delclaux et al., 2008) using the GR_B + THMB Model, includes precipitation and reference evapotranspiration as inputs, which are adjusted by a coefficient C, set such that the Nash coefficient is maximized for the monthly flows of the Chari-Logone River system. Precipitation is then split between a soil reservoir with maximum capacity A, and surface runoff, the amounts depending on the level of water in the soil reservoir. The soil reservoir drains through actual evapotranspiration and percolation, which generates sub-surface runoff. Elevation data from the Shuttle Radar Topography Mission (SRTM) are used for generating drainage directions and water accumulation areas. Two irrigation scenarios were modeled. Lake level data was used to validate the model results. In comparison, our model also uses precipitation and actual evapotranspiration as inputs, though only for the wet season. The timing of subsurface flow and surface flow through the river systems is simulated in our model by the delay between the wet season end and the month being forecast, and by the HydroBASINS shapefile for the southern Lake Chad Basin, which defines the limits of our representation of the lake watershed. Lake level is used in the GR_B + THMB model as memory of previous conditions, whereas the lake area for the previous year is used in our model for this purpose. Unlike the GR_B + THMB model, we do not model irrigation withdrawals; as discussed in the Introduction section, this is a small part of the overall hydrology of the Lake Chad Basin. While the GR_B + THMB model uses stream flow data to calibrate the model, and lake elevation to validate the model, we use lake area estimates for these tasks and lake elevation as an additional input parameter.

The main limitation of the model we present here is that it is not a physically-based model and may not perform well for conditions outside of those in the database we have developed. For example, we are not able to do regressions and provide predictions for wet season months because we do not have surface water area data for those conditions (Policelli et al., 2018). Also, the performance may degrade if it is used outside of the range of areas for which the regression models were established (8,700 sq. km – 16,800 sq. km) or when the hydrology of the lake changes substantially, such as when “small Lake Chad” transitions to “normal Lake Chad” at about 18,000 sq. km. There is no fixed date at which the model becomes unusable. However, if it is used for operational forecasting, it would be wise to regularly update the model with new data as it becomes available.

6. Conclusions

We have built a record of remote sensing data and model products (precipitation, evapotranspiration, lake height, and lake area) for the Lake Chad Basin and used this record to run a correlation analysis and a regression analysis. From the correlation analysis (see Appendix), we have found (1) the highest correlation between basin evaporation and total lake surface water area is 0.43 and occurs with a seven month lag, (2) the highest correlation between lake height anomaly and total lake surface water area is 0.57 and occurs with a four month lag, (3) the highest correlation between precipitation and total lake surface water area is 0.39 and occurs with a seven month lag, (4) the highest correlation between percent of 1988-2016 average lake ET and the lake total surface water area is 0.65 at zero lag time, and (5) there is a correlation of 0.63 between the surface water area and the previous year's surface water area for the same month.

Note that lake surface height is more closely correlated with total lake surface water area than is precipitation, as might be expected from a measurement further "downstream". Additionally, basin ET is closely correlated with basin precipitation as might be expected since the precipitation is the source water for ET. Finally, basin ET is slightly more closely correlated with total surface water area than basin precipitation.

From our regression analysis, we have derived a set of equations that can be used starting in late November for predicting the total surface water area of Lake Chad for a given month (except November) during the dry season. The best of these in terms of R-squared use all of our parameters: wet season precipitation (Pws), wet season south basin evapotranspiration (ETws), wet season lake evapotranspiration percent of 1988-2016 average (LakeETws), lake height variation relative to 2002-2009 (H), and the previous year's surface area for the given month (A), though we are likely overfitting the data when using all of these variables. The results of Leave One Out Cross Validation (LOOCV) testing and backward elimination variable selection show the best performing variables for December to be LakeETws, Pws and A-, for January they are ETws, H, and A-, and for February, March, April and May they are Pws, H, and A-. The regression equations perform at between 5.3 and 7.6 average absolute percent error in the LOOCV testing, and outperform predictions made using the average value for the given month or the previous year's total surface water area.

The predictions using all of the equations derived from regression in this study can be made with only remotely sensed data and model outputs; no in-situ data is required. Any improvements in the measurement of the parameters we use in this analysis would likely improve the desired end result – the prediction of the Lake Chad surface area in time to be used for agricultural decisions.

Acknowledgements

Support for this research was provided by the NASA Goddard Space Flight Center's Academic Investment for Mission Success (AIMS) Program. Support for this research was also provided by NASA under its Research Opportunities in Space and Earth Sciences (ROSES)-2013 Interdisciplinary Studies (IDS) Program (solicitation NNH13ZDA001N-IDS) through the Radiation Sciences Program. We appreciate the efforts of providers of the data products used for this study from various satellite sensors and models, particularly lake imagery from USGS, lake height data from USDA, CHIRPS precipitation data from U.C. Santa Barbara, evapotranspiration data from the FEWS NET Land Data Assimilation System (FLDAS), the southern Lake Chad Basin shapefile from WWF, meteorological data from NOAA NCEP-DOE, and Lake Chad surface water area from (Leblanc et al., 2011).

Role of the funding Sources: The funding sources had no role in the study design; in the collection, analysis and interpretation of data; in the writing of the report; or in the decision to submit the article for publication.

Author Contributions: Frederick Policelli and Ben Zaitchik conceived and designed the experiments; Frederick Policelli and Alfred Hubbard performed the experiments; Frederick Policelli, Hahn Chul Jung, and Ben Zaitchik analyzed the data; Charles Ichoku helped conceive the experiments and provided domain knowledge; Frederick Policelli wrote the paper with contributions from all co-authors.

Declarations of Interest: None

References

Austin, P.C., Steyerberg, E.W., 2015. The number of subjects per variable required in linear regression analyses. J. Clin. Epidemiol. 68, 627–636. https://doi.org/10.1016/j.jclinepi.2014.12.014

Bader, J.-C., Lemoalle, J., Leblanc, M., 2011. Modèle hydrologique du lac Tchad. Hydrol. Sci. Journal–Journal Sci. Hydrol. 56, 411–425.

Badr, H.S., Dezfuli, A.K., Zaitchik, B.F., Peters-Lidard, C.D., 2016. Regionalizing Africa: Patterns of Precipitation Variability in Observations and Global Climate Models. J. Clim. 29, 9027–9043. https://doi.org/10.1175/JCLI-D-16-0182.1

Birkett, C.M., 2000. Synergistic remote sensing of Lake Chad: Variability of basin inundation.

Remote Sens. Environ. 72, 218–236.

Busker, T., de Roo, A., Gelati, E., Schwatke, C., Adamovic, M., Bisselink, B., Pekel, J.-F., Cottam, A., 2018. A global lake and reservoir volume analysis using a surface water dataset and satellite altimetry. Hydrol. Earth Syst. Sci. Discuss. 1–32. https://doi.org/10.5194/hess-2018-21

Carmouze, J.-P., Durant, J.R., Lévêque, C., 1983. Lake Chad: ecology and productivity of a shallow tropical ecosystem. Springer Science & Business Media.

Coe, M.T., Birkett, C.M., 2004. Calculation of river discharge and prediction of lake height from satellite radar altimetry: Example for the Lake Chad basin. Water Resour. Res. 40, W10205. https://doi.org/10.1029/2003WR002543

Coe, M.T., Foley, J.A., 2001. Human and natural impacts on the water resources of the Lake Chad basin. J. Geophys. Res. Atmospheres 106, 3349–3356. https://doi.org/10.1029/2000JD900587

Delclaux, F., Le Coz, M., Coe, M., Favreau, G., Ngounou Gatcha, B., 2008. Confronting models with observations for evaluating hydrological change in the lake chad basin, Africa, in: XIIIth World Water Congress. URL: https://www.iwra.org/member/congress/resource/abs355_article.pdf accessed 7/18/2018

Funk, C., Peterson, P., Landsfeld, M., Pedreros, D., Verdin, J., Shukla, S., Husak, G., Rowland, J., Harrison, L., Hoell, A., Michaelsen, J., 2015. The climate hazards infrared precipitation with stations—a new environmental record for monitoring extremes. Sci. Data 2, sdata201566. https://doi.org/10.1038/sdata.2015.66

Gao, H., Bohn, T.J., Podest, E., McDonald, K.C., Lettenmaier, D.P., 2011. On the causes of the shrinking of Lake Chad. Environ. Res. Lett. 6, 034021.

Grove, A.T., 1996. African river discharges and lake levels in the twentieth century. Limnol. Climatol. Paleoclimatology East Afr. Lakes Gordon Breach Neth. 95–100.

Harrell, F.E., Lee, K.L., Mark, D.B., 1996 Multivariable Prognostic Models: Issues in Developing Models, Evaluating Assumptions and Adequacy, and Measuring and Reducing Errors. Stat. Med. 15, 361–387. https://doi.org/10.1002/(SICI)1097-0258(19960229)15:4<361::AID-SIM168>3.0.CO;2-4

Lake Chad flooded savanna", World Wildlife Fund, no date, https://www.worldwildlife.org/ecoregions/at0904 accessed 7/18/2018

Leblanc, M., Lemoalle, J., Bader, J.-C., Tweed, S., Mofor, L., 2011. Thermal remote sensing of water under flooded vegetation: New observations of inundation patterns for the ‘Small’ Lake Chad. J. Hydrol. 404, 87–98. https://doi.org/10.1016/j.jhydrol.2011.04.023

Lehner, B., Grill G., 2013. Global river hydrography and network routing: baseline data and new approaches to study the world's large river systems. Hydrological Processes, 27(15): 2171-2186.

Lehner, B., Verdin, K., Jarvis, A., 2011 New Global Hydrography Derived From Spaceborne Elevation Data. Eos Trans. Am. Geophys. Union 89, 93–94. https://doi.org/10.1029/2008EO100001

Lemoalle, J., 2004. Lake Chad: A Changing Environment. In: Nihoul J.C.J., Zavialov P.O., Micklin P.P. (eds) Dying and Dead Seas Climatic Versus Anthropic Causes. NATO Science Series: IV: Earth and Environmental Sciences, vol 36. Springer, Dordrecht

Lemoalle, J., Bader, J.-C., Leblanc, M., Sedick, A., 2012. Recent changes in Lake Chad: Observations, simulations and management options (1973–2011). Glob. Planet. Change 80, 247–254.

McNally, A., Arsenault, K., Kumar, S., Shukla, S., Peterson, P., Wang, S., Funk, C., Peters-Lidard, C.D., Verdin, J.P., 2017. A land data assimilation system for sub-Saharan Africa food and water security applications. Sci. Data 4, 170012. https://doi.org/10.1038/sdata.2017.12

Morton, F.I., 1986. Practical Estimates of Lake Evaporation. J. Clim. Appl. Meteorol. 25, 371–387. https://doi.org/10.1175/1520-0450(1986)025<0371:PEOLE>2.0.CO;2

Okpara, U.T., Stringer, L.C., Dougill, A.J., 2016. Lake drying and livelihood dynamics in Lake Chad: Unravelling the mechanisms, contexts and responses. Ambio 45, 781–795. https://doi.org/10.1007/s13280-016-0805-6

Policelli, F., Hubbard, A., Jung, H.C., Zaitchik, B., Ichoku, C., 2018. Lake Chad Total Surface Water Area as Derived from Land Surface Temperature and Radar Remote Sensing Data. Remote Sens. 10, 252. https://doi.org/10.3390/rs10020252

Ricko, M., Carton, J.A., Birkett, C.M., Cretaux, J.-F., 2012. Intercomparison and validation of continental water level products derived from satellite radar altimetry. J. Appl. Remote Sens. 6, 061710. https://doi.org/10.1117/1.JRS.6.061710

Sarch, M.-T., Birkett, C., 2000. Fishing and farming at Lake Chad: Responses to lake-level fluctuations. Geogr. J. 166, 156–172.

Satellite Radar Altimetry: Global Reservoir and Lake Elevation Database [WWW Document], no date URL https://www.pecad.fas.usda.gov/cropexplorer/global_reservoir/ (accessed 10.17.17).

Schneider, Udo; Becker, Andreas; Finger, Peter; Meyer-Christoffer, Anja; Rudolf, Bruno; Ziese, Markus (2015): GPCC Full Data Monthly Product Version 7.0 at 0.5^0 : Monthly Land-Surface Precipitation from Rain-Gauges built on GTS-based and Historic Data. DOI: 10.5676/DWD GPCC/FD M V7 050

Senay, G.B., Bohms, S., Singh, R.K., Gowda, P.H., Velpuri, N.M., Alemu, H., Verdin, J.P., 2013 Operational Evapotranspiration Mapping Using Remote Sensing and Weather Datasets: A New Parameterization for the SSEB Approach. JAWRA J. Am. Water Resour. Assoc. 49, 577–591. https://doi.org/10.1111/jawr.12057

Zhu, W., Yan, J., Jia, S., 2017a. Monitoring Recent Fluctuations of the Southern Pool of Lake Chad Using Multiple Remote Sensing Data: Implications for Water Balance Analysis. Remote Sens. 9, 1032. https://doi.org/10.3390/rs9101032

Appendix

We examined the correlation between P–ET for the southern Lake Chad Basin and the lake elevation variation relative to 1993-2002 for 1992 to 2017, and found the maximum correlation of 0.69 at four months lag time between these variables. Next, we examined the correlation between the P–ET for the southern part of the basin and the lake surface area and found the maximum correlation of 0.37 at eight months lag time. We also examined the correlation between the lake elevation variation and the lake surface water area and found the maximum correlation of 0.57 at four months lag time. We found a maximum correlation of 0.43 for ET vs the lake's surface water area at 7 months, and a correlation of 0.63 for the surface water area versus the previous year's surface water area for the same month. There is apparently memory of the previous year's area in the system.

To determine the value added by the ET data to our analysis, we examined the correlation between precipitation (without subtracting ET) for the southern Lake Chad Basin and the lake elevation and found an increase in the maximum correlation to 0.80 at four months lag time. We found the correlation between the percentage of the 1988-2016 average lake ET and the lake total surface water area to be 0.65 at zero lag time. We also examined the correlation between precipitation and total lake surface area and found an increase in the maximum correlation to 0.39 at seven months lag time. This is the same lag time as for the maximum correlation between ET and total lake surface area and represents the time it takes for much of the net precipitation to make its way from run off in the southernmost part of the Lake Chad basin, to flowing through the Chari-Logone River system, to reaching the lake and causing an increase in the lake area. A surprising result of the correlation analysis is that the use of FLDAS ET data in the analysis to produce (P-ET) causes a small decrease in the correlation numbers relative to what is achieved with precipitation alone. Note however that ET is somewhat more closely correlated with total lake surface water area than is precipitation (.43 vs. .39, both at 7 months lag time). Precipitation and ET have a maximum correlation coefficient of 0.93 with a one month delay between the two.