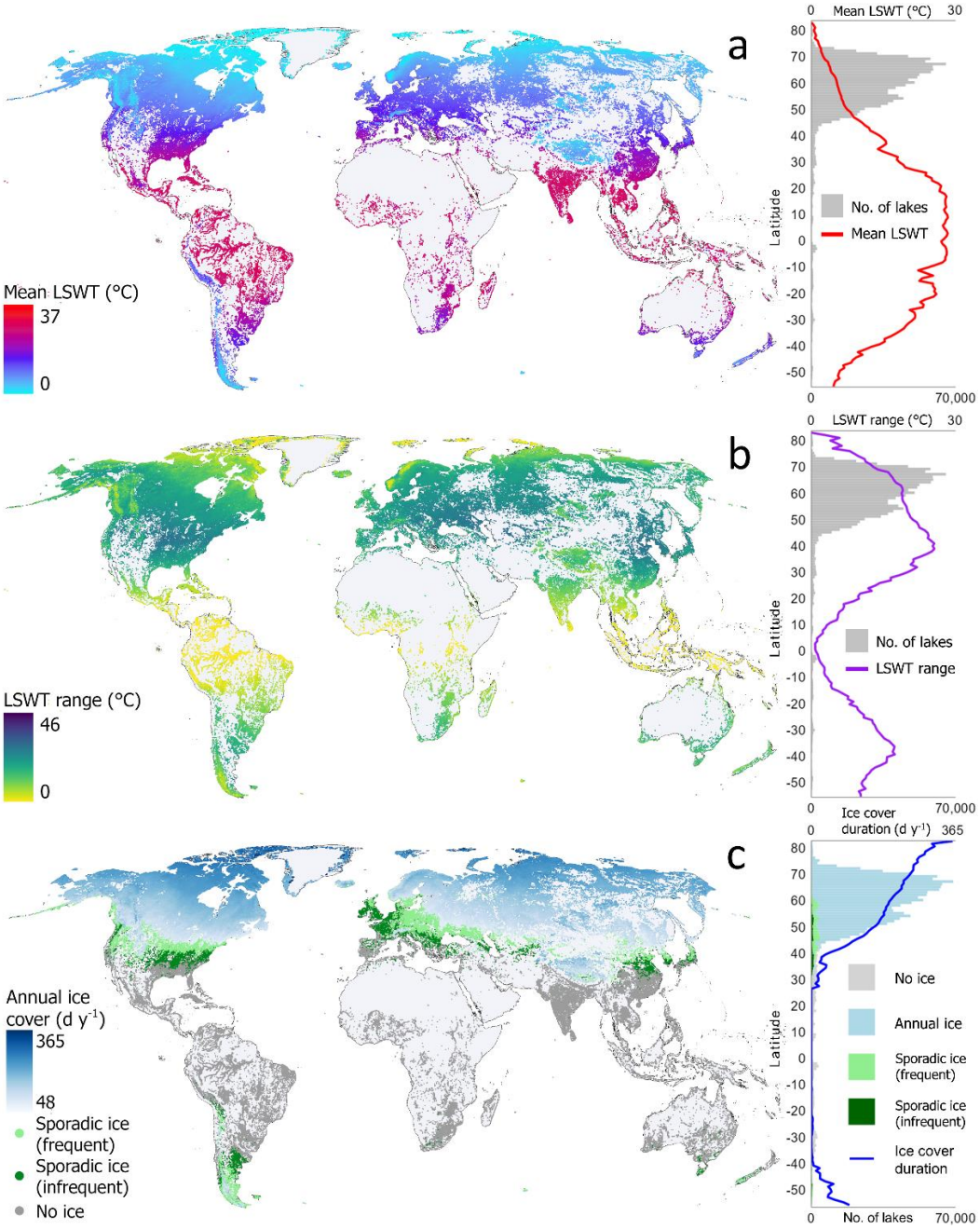
LakeTEMP

Technical Documentation – version 1.0

Prepared by Maartje Korver (maartje.korver@mail.mcgill.ca) and Bernhard Lehner (bernhard.lehner@mcgill.ca)



1. Introduction

This documentation describes the technical aspects of the LakeTEMP dataset, consisting of two separate subsets termed “Primary” and “Aggregated”, each stored in separate .csv data files. The Primary dataset consists of quality-controlled lake surface water temperature (LSWT) observations derived from satellite imagery (Landsat 8) for ~1.4 million lakes ≥ 0.1 km^2, and the Aggregated dataset consists of monthly and yearly summary statistics, including estimations of ice phenology, interpolated from the Primary dataset (Fig. 1).

The development of LakeTEMP is fully described in Korver et al. (2024) and should be cited as:

Korver, M. C., Lehner, B., Cardille, J. A., & Carrea, L. (2024). Surface water temperature observations and ice phenology estimations for 1.4 million lakes globally. Remote Sensing of Environment, 308, 114164. https://doi.org/10.1016/j.rse.2024.114164

The following sections describe the general methods and data characteristics, provide data format explanations, and state the license, disclaimer, and acknowledgements associated to the data.

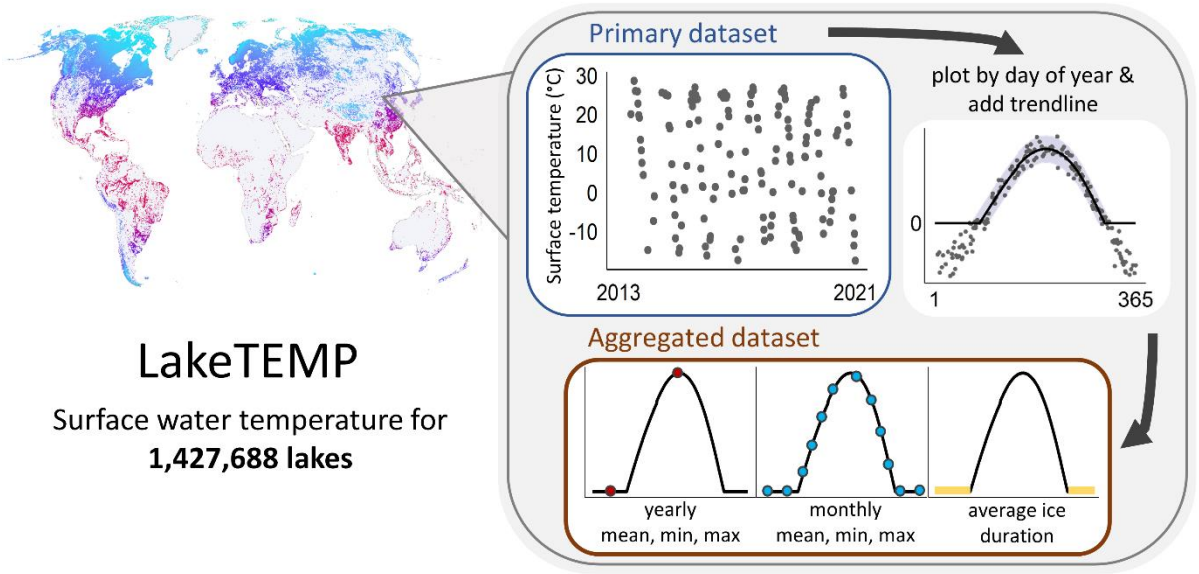


Figure 1: Overview of the LakeTEMP dataset and its two subsets. The Primary dataset comprises weekly to monthly discrete LSWT data for the period 2013 - 2021. The Aggregated dataset comprises average yearly and average monthly mean, minimum, and maximum LSWTs (± 95 % confidence intervals), and ice phenology predictions (average yearly duration of ice cover and the approximate average ice on and ice off dates; ± 95 % confidence intervals), derived from the Primary data using seasonal trendlines.

2. Methods and data characteristics

A detailed description of the methodologies used to develop LakeTEMP is provided by Korver et al. (2024).

The source of original lake locations was provided by the HydroLAKES database (Messager et al., 2016), which includes shorelines of natural (1,420,891) as well as artificial (6,797) lakes of at least 0.1 km^2 in surface area between 81° N and 56° S. For more information on HydroLAKES please refer to the Technical Documentation at www.hydrosheds.org/hydrolakes.

The Primary dataset consists of LSWT observations with quality flags between 2013 – 2021 for each of the 1,427,688 lakes. The data was derived from Landsat 8 thermal infrared imagery (USGS, 2021), acquired from a 50 x 50 m area around the center point location of each lake, which is the single location inside the lake that is the furthest away from any lake shore or island. The temporal resolution of the dataset is generally around 16 days, but it can range from multiple readings a week to only a few observations per year depending on latitude (shorter Landsat revisit intervals at higher latitudes), cloud conditions (cloud contaminated readings are removed) and potential overlap between orbital paths (increasing the revisit interval).

The Aggregated dataset consists of the average yearly and monthly mean (± 95 % confidence intervals), maximum, and minimum LSWTs, the average yearly lake ice cover durations (± 95 % confidence intervals), and the approximate average ice on and ice off dates. These statistics were calculated from the Primary data by fitting a seasonal trendline through each LSWT timeseries (Fig. 1).

Through the use of the unique HydroLAKES identifiers, this dataset can be easily combined with other global lake datasets, such as the LakeATLAS database (Lehner et al., 2022), which provides 56 hydro-environmental characteristics for the same ~1.4 million lakes and their lake watersheds. For more information on LakeATLAS please refer to the documentation of the overarching HydroATLAS product at www.hydrosheds.org/hydroatlas.

3. Data format and availability

Both datasets are provided as single CSV files and can be downloaded from https://figshare.com/s/f05ffaae54c2215243c6

3.1 Primary dataset

Table 1 displays the name and explanation of each variable in the Primary dataset.

The quality of each individual LSWT observation is indicated with a 'water' and an 'outlier' flag. The water flag refers to the occurrence of water at the time and location of each observation, derived from the Global Surface Water dataset (Pekel et al., 2016), with flags indicating water (1) no water or 'land' (2), unknown surface (3), or no information (4). The outlier flag indicates whether the observation is within accepted range (0), is an extreme value, but potentially a valid observation (1), or is an outlier (2). Please note that no (valid) observations were available for 67 lakes (including the Caspian Sea) due to consistent cloud cover and for 4,376 lakes due to consistently dry surface conditions.

For high-quality data, we advice not to use the observations taken from dry surfaces (water = 2) or those flagged as outliers (outlier = 2). Note that observations < 0 °C were not removed from the dataset because they were used for the estimation of ice phenology in the Aggregated data. However, these values should not be interpreted as ice surface temperatures, which require a different, locally calibrated algorithm for their calculation.

Table 1: The variable names and explanations of the Primary data.

Variable	Explanation
Hylak_id	Unique identifier of lake, from HydroLAKES (Messager et al., 2016)
dateTime	Date and time of measurement (yyyy-mm-dd hh:mm:ss)
temp	Lake surface water temperature (°C * 100)
water	Flag to indicate whether observation was taken from 'water' (1), 'land' (2), 'unknown' surface (3), or if no information was available (4).
outlier	Flag to indicate if observation is not an outlier, i.e., within accepted range (0), is an extreme value, but potentially a valid observation (1), or is an outlier (2).

3.2 Aggregated dataset

Table 2 displays the name and explanation of each variable in the Aggregated dataset.

Lake metadata (name, country, continent, lake type, area, depth, and elevation) were collected from the HydroLAKES dataset. For further technical details of these variables consult the Technical Documentation at www.hydrosheds.org/hydrolakes.

For each lake, the overall data quality is indicated by the method used for statistical calculations (i.e., from trendline or direct, with higher data quality for trendline derived statistics), and the likelihood that a lake experienced intermittency. Table 3 outlines in more detail how these flags were calculated, how they can be interpreted, what the consequences are for the data, and how many lakes in the dataset were affected. In addition, data quality is affected by the number of observations used for statistical calculations ('n_obs') and as a guideline, data quality can be considered good when n_obs > 85, acceptable when n_obs is between 46 and 85, and minimal when n_obs is between 12 and 45.

For high-quality data, we advice not to use statistics that were calculated directly (*stat_method* = 1), were calculated from lakes with potential intermittency (*intermittency* = 1, 2, 3, or 4), or that were calculated from less than 46 observations.

Table 2: The variable names and explanations of the Aggregated dataset.

Variable	Explanation
Hylak_id	Unique identifier of lake, from HydroLAKES (Messager et al., 2016)
center_long	Longitude of LSWT acquisition location (lake center point) (decimal degrees)
center_lat	Latitude of LSWT acquisition location (lake center point) (decimal degrees)
Lake_name	Name of lake or reservoir. This field is currently only populated for lakes with an area of at least 500 km^2; for large reservoirs where a name was available in the GRanD (Lehner et al., 2011) database; and for smaller lakes where a name was available in the GLWD (Lehner & Doll, 2004) database.
Country	Country that the lake (or reservoir) is located in.
Continent	Continent that the lake (or reservoir) is located in. Geographic continent: Africa, Asia, Europe, North America, South America, or Oceania (Australia and Pacific Islands)
Lake_type	Indicator for lake type: 1: Lake, 2: Reservoir, 3: Lake control (i.e. natural lake with regulation structure).
Lake_area	Lake surface area (i.e., polygon area) (km^2)
Depth_avg	Average lake depth (m)
Elevation	Elevation of lake surface (m above sea level)
n_obs	Number of LSWT observations used to calculate statistics
stat_method	Flag to indicate calculation method. 0: calculated from a seasonal trendline. 1: calculated from observations directly.
intermittency	Prior to observation quality control and filtering, the timeseries consisted of 0: < 20 % 'land' over 'water' observations, 1: > 20 % and < 80 % 'land' over 'water' observations, 2: > 80 % 'land' over 'water' observations, 3: 100 % a combination of 'land' and 'unknown' observations, 4: 100 % 'land' observations, or 5: 100 % readings with no information on water occurrences.
Tyear_mean	The average yearly mean LSWT, assuming 0 °C during periods of ice cover (°C * 100)
Tyear_mean_open	The average yearly mean LSWT from open water observations only (°C * 100)
Tyear_min	The average yearly minimum LSWT, i.e., the minimum of 365 daily trendline values (°C * 100)
Tyear_max	The average yearly maximum LSWT, i.e., the maximum of 365 daily trendline values (°C * 100)
Tyear_CImean_min	97.5 % likelihood that the average yearly mean LSWT is above this value (°C * 100)
Tyear_CImean_max	97.5 % likelihood that the average yearly mean LSWT is below this value (°C * 100)
ice_days	The average yearly duration of ice cover (days)
ice_days_min	Lower estimate of the average yearly duration of ice cover, based on CIpred (95% likelihood of LSWT measurement range) (days)
ice_days_max	Upper estimate of the average yearly duration of ice cover, based on CIpred (95% likelihood of LSWT measurement range) (days)
ice_start	The average yearly start of ice on (Julian day)

ice_start_min	Lower estimate of the average yearly start of ice on, based on CIpred (95% likelihood of LSWT measurement range) (Julian day)
ice_start_max	Upper estimate of the average yearly start of ice on, based on CIpred (95% likelihood of LSWT measurement range) (Julian day)
ice_end	The average yearly end of ice off (Julian day)
ice_end_min	Lower estimate of the average yearly end of ice off, based on CIpred (95% likelihood of LSWT measurement range) (Julian day)
ice_end_max	Upper estimate of the average yearly end of ice off, based on CIpred (95% likelihood of LSWT measurement range) (Julian day)
Tmonth_mean_x	The average mean LSWT of month x (°C * 100)
Tmonth_min_x	The average minimum LSWT of month x, i.e., the minimum of 28 to 31 daily trendline values (°C * 100)
Tmonth_max_x	The average maximum LSWT of month x, i.e., the maximum of 28 to 31 daily trendline values (°C * 100)
Tmonth_CImean_min_x	97.5 % likelihood that the average monthly mean LSWT is above this value (°C * 100)
Tmonth_CImean_max_x	97.5 % likelihood that the average monthly mean LSWT is below this value (°C * 100)
Tmonth_CIpred_min_x	97.5 % likelihood that the range of LSWT measurements in month x is above this value (°C * 100)
Tmonth_CIpred_max_x	97.5 % likelihood that the range of LSWT measurements in month x is below this value (°C * 100)

Table 3: Flagging scheme of the Aggregated dataset, with flag definitions based on the underlying Primary dataset, interpretations of the lake types associated to each flag, consequences for the data, and the % of lakes affected (out of 1,427,688). The definitions of the 'intermittency' flags are given by the % of timeseries observations that were acquired from 'water', 'land', 'unknown', or 'no information' surfaces, as derived from Pekel et al. (2016). Interpretations are descriptions of the likely condition of a lake during data acquisition, however other interpretations are possible.

Flag	Definition	Interpretation	Consequences	% of lakes
Calculation method				
0	≥ 12 observations	All regions/lake types	Full set of statistics calculated from trendline	98.8
1	< 12 observations	Lake in high cloud-cover area or intermittent/ephemeral lake	Only yearly mean, min, max calculated directly	0.8
NA	No observations	No observations due to permanent cloud cover or intermittency flag 4	No statistics calculated	0.4
Intermittency				
0	≤ 20 % 'land' over 'water' observations	Perennial lake, or dry for periods that are not long enough to affect calculations considerably	'Land' observations removed before calculation	87.2
1	> 20 % and ≤ 80 % 'land' over 'water' observations	Intermittent lake	'Land' observations removed before calculation	6.2
2	> 80 % 'land' over 'water' observations	Ephemeral lake or intermittent lake with long dry season	'Land' observations removed before calculation	2.7
3	100 % 'land' + 'unknown' observations	Permanently dry lake with seasonal snow or ice cover	Statistics calculated from 'unknown' readings only	3.4
4	100 % 'land' observations	Permanently dry lake; or wetland with dense vegetation	No statistics calculated	0.3
5	100 % 'no information'	Lake near north pole; or lake in area with high cloud cover; or other unknown issues	Statistics calculated from all observations	0.2

4. License, disclaimer, and acknowledgements

General citations and acknowledgements of these datasets should be made as follows:

Korver, M. C., Lehner, B., Cardille, J. A., & Carrea, L. (2024). Surface water temperature observations and ice phenology estimations for 1.4 million lakes globally. Remote Sensing of Environment, 308, 114164. https://doi.org/10.1016/j.rse.2024.114164

5. References

Lehner, B., & Doll, P. (2004). Development and validation of a global database of lakes, reservoirs and wetlands. *Journal of Hydrology*, 22.

Lehner, B., Liermann, C. R., Revenga, C., Vörösmarty, C., Fekete, B., Crouzet, P., Döll, P., Endejan, M., Frenken, K., Magome, J., Nilsson, C., Robertson, J. C., Rödel, R., Sindorf, N., & Wisser, D. (2011). High-resolution mapping of the world's reservoirs and dams for sustainable river-flow management. *Frontiers in Ecology and the Environment*, *9*(9), 494–502. https://doi.org/10.1890/100125

Lehner, B., Messager, M. L., Korver, M. C., & Linke, S. (2022). Global hydro-environmental lake characteristics at high spatial resolution. *Scientific Data*, *9*(1), Article 1. https://doi.org/10.1038/s41597-022-01425-z

Messager, M. L., Lehner, B., Grill, G., Nedeva, I., & Schmitt, O. (2016). Estimating the volume and age of water stored in global lakes using a geo-statistical approach. *Nature Communications*, *7*(1), 13603. https://doi.org/10.1038/ncomms13603

Pekel, J.-F., Cottam, A., Gorelick, N., & Belward, A. S. (2016). High-resolution mapping of global surface water and its long-term changes. *Nature*, *540*(7633), 418–422. https://doi.org/10.1038/nature20584

United States Geological Survey. (2021). Landsat 8 Collection 1 Tier 1 TOA Reflectance. Accessed from Google Earth Engine between August 9 – August 21, 2021.