OSEANOLOGI DAN LIMNOLOGI DI INDONESIA
Online ISSN: 2477-328X
Akreditasi RISTEKDIKTI No. 200/M/KPT/2020
http://oldi.lipi.go.id

A Simple Solution for Refining Lake Water Temperature Profiles Data Arrayed from High-Frequency Monitoring Sensors

Arianto B. Santoso[1]*, Endra Triwisesa[1], Muh. Fakhrudin[1]

Research Centre for Limnology, Indonesian Institute of Sciences
Jalan Raya Jakarta-Bogor Km. 46 Cibinong, Bogor 16911

*Email: ari@limnologi.lipi.go.id

Submitted 10 June 2020. Reviewed 10 August 2020. Accepted 17 November 2020.
DOI: 10.14203/oldi.2021.v6i1.324

Abstract

The revolutionized aquatic monitoring sensors are essential in capturing environmental patterns that traditional discrete samplings might not be able to. They allow scientists to further synthesize and better conclude processes in aquatic ecosystems. These sensors produce high-frequency data that provide information on a fine temporal scale, even near real-time. The massive quantities of the streamed data, however, create challenges for scientists to grasp the concrete information. Filtering data quality, on the other hand, is another problem scientists might have encountered as sensor accuracy and precision may drift along the line. Hence, quality assurance and quality control might be quite labouring owing to the size of datasets to handle. This paper proposed a semi-mechanistic algorithm to improved false water temperature data. Using "theoretical" thermal stratification as a reference, this algorithm fixed sensors error readings. A 5-month dataset of water temperature profiles of Lake Maninjau, West Sumatra, captured every 10 minutes from a set of sensors in thermistor chain was applied. We found that most data fit to the theoretical temperature profile, R^2 = 0.962, RMSE = 0.081°C. A number of errors, however, were observed in the upper layer of the lake (<20 m), the most dynamic layer in terms of its thermal variation. Sensor drifts in this active upper mixed layer can be related to the generated errors. Through this simple solution, not only improving the quality of the observed water temperature data, but was also able to identify the most probable source of errors.

Keywords: quality assurance/quality control, monitoring sensor, water temperature, lakes

Abstrak

Solusi Sederhana untuk Menyaring Data Profil Suhu Air Danau dari Sensor Pemantauan Frekuensi Tinggi. Dalam pemantauan lingkungan akuatik, keberadaan sensor sangatlah krusial untuk menangkap fenomena alam yang belum tentu dapat dilakukan oleh sampling secara manual. Dengan adanya sensor pemantau, analisa dan sintesa proses-proses pada ekosistem akuatik akan jauh lebih baik. Sistem pemantau tersebut mampu menghasilkan data dengan resolusi waktu yang tinggi, bahkan mendekati "*real time*". Namun demikian, dengan masifnya jumlah data yang dihasilkan, penanganan dan kontrol kualitas data merupakan suatu keniscayaan dalam menyimpulkan informasi. Untuk memberikan solusi mudah dalam memperbaiki pembacaan data suhu profil air, dalam kajian ini digunakan algoritma semi mekanistik. Dengan menggunakan "teori" stratifikasi suhu sebagai acuan, algoritma ini memperbaiki kesalahan pembacaan sensor. Pengujian dalam penggunaan algoritma dilakukan pada rangkaian data suhu air Danau Maninjau, Sumatra Barat, selama lima bulan dengan observasi tiap 10 menit menggunakan sensor suhu yang terangkai di stasiun pengamatan. Secara umum data hasil pengukuran sesuai dengan teori, yaitu R^2 = 0.962, RMSE = 0.081^{o}C. Simpangan dalam pembacaan data sebagian besar terdeteksi pada lapisan permukaan air danau (<20 m) yang merupakan lapisan paling aktif dalam hal dinamika suhunya. Kajian ini menunjukkan bahwa penggunaan algoritma semi mekanistik merupakan solusi mudah dalam penanganan dan kontrol kualitas data, baik dalam hal perbaikan data maupun identifikasi kesalahan pada sensor.

Kata kunci: kontrol kualitas data, monitoring sensor, suhu air, danau

Introduction

As sentinels of climate change and anthropogenic impacts, lakes are now being assessed and monitored by society around the globe (Adrian et al., 2009; Hanson et al., 2016). This aims to better understand how lakes behave in response to anthropogenic force and climate changes (Hamilton et al., 2014). With the advances in sensor technology information gained from such a monitoring program are more robust and may provide opportunities to observe many ecological variables at finer temporal scales not possible with traditional sampling techniques. Our understanding of lake metabolism and air-water gas transfer (Dugan et al., 2016), hydrodynamics (Hipsey et al., 2019), lake physics and episodic events (Andersen et al., 2020), has been supported by the high-frequency data from those sensors. Lake managers might also take advantage of this technology and leverage the output as information to predict possible scenarios in restoration program (Hipsey et al., 2015).

Aquatic monitoring sensors might have been revolutionized our understanding of lake dynamics. High-frequency data generated by those sensors might inform scientists and managers about phenomena occurring in the lake. This high-frequency resolution (i.e. hourly or minute bases) creates packages of big data (Porter et al., 2012) which, however, can not escape from errors. Sensor malfunctions are susceptible and they can result in lost or poor-quality data. Filtering the incoming data, therefore, is a must task. As the quantity of the data is quite massive, quality assurance and control (QA/QC) might have been a labouring task, although such evaluation is crucial in minimizing faulty results.

This study proposed a semi-mechanistic algorithm to improve the quality of high-frequency water temperature data. Using thermal stratification as reference (see Wetzel, 2001), the algorithm fixed sensors error readings. This algorithm was applied to a 5-month data set of water temperature profiles of Lake Maninjau, West Sumatra, captured every 10 minutes from a set of sensors in thermistor chain. The possible source of errors that cause sensor faulty readings was also considered. The importance of high-frequency monitoring systems in tropical deep lakes of Indonesia to better understand their dynamics was also emphasized, knowing that water temperature might vary on a sub-daily scale (Katsev et al., 2010; Santoso et al., 2018).

Methods

A time-series water temperature data of Maninjau were evaluated, a 165 m depth tropical volcanic lake in West Sumatra (Figure 1). Data was arrayed from a monitoring buoy, the On Line Monitoring System of Research Centre for Limnology-LIPI (known as E-MOST), deployed in the lake. The E-MOST was equipped with a set of Dallas DS18B20 temperature sensors placed between 0.5 m to 62 m depth with an interval of 2 m. Packages of data were sent from E-MOST to the data centre in Cibinong, Bogor, through a GPRS network (Figure 2). A 5-month data package, from 1 January 2018 to 28 May 2018, recorded every 10 minutes, was employed in this evaluation.

A semi-mechanistic algorithm (Minns & Shutter, 2013; Mackenzie-Grieve & Post, 2006) was applied to seek the typical shape of water temperature profiles in which epilimnion, thermocline, and hypolimnion layers are apparent (Figure 3). Theoretically, most lakes, particularly in temperate zones, are seasonally stratified into those layers (Wetzel, 2001). Deep tropical lakes, however, are also stratified into those thermal layers, although the temperature difference between surface and bottom water is relatively small, ~3.5^{o}C (Katsev et al., 2010; Santoso et al., 2018). The shape of each temperature profile was described by:

$$T_{(z)} = T_{hyp} + (T_{epi} - T_{hyp}) * \frac{Z_{mix}{}^{S}}{Z_{mix}{}^{S} + Z^{S}}$$

where $T_{(z)}$ is the water temperature (°C) at a given depth Z (m), T_{hyp} is the hypolimnion temperature, T_{epi} is the epilimnion temperature, Z_{mix} is the assumed thermocline depth, and S (dimensionless) is the steepness transition depth from epilimnion to hypolimnion. To assist the estimation of thermal layers from the observation data (see theoretical basis given by Read et al., 2011) an R package rLakeAnalyzer was used (Winslow et al., 2015) to estimate the metalimnion depth which is the water stratum of steep thermal gradient demarcated by the intersections of the nearby homoiothermal epilimnion and the hypolimnion (Figure 3). Epilimnion, then, was described as the layer above upper metalimnion, while hypolimnion lays below lower metalimnion. T_{hyp} and T_{epi} were set as the volumetrically averaged temperature of hypolimnion and epilimnion, respectively. An optimization routine ("L-BFGS-B": R Core Team, 2019) was used to derive Z_{mix} and S by minimizing the negative log-likelihood of the errors between the new shape temperature profile and data observation. Hence, through the algorithm above, "theoretical" temperature profiles were regenerated based on sensor observations. Probable error in sensors was identified by the magnitude of root mean square error (RMSE) of raw data deviated from the optimized temperature profile. All computation and analyses were performed in the R statistical package (R Core Team, 2019; Version 3.5.3).

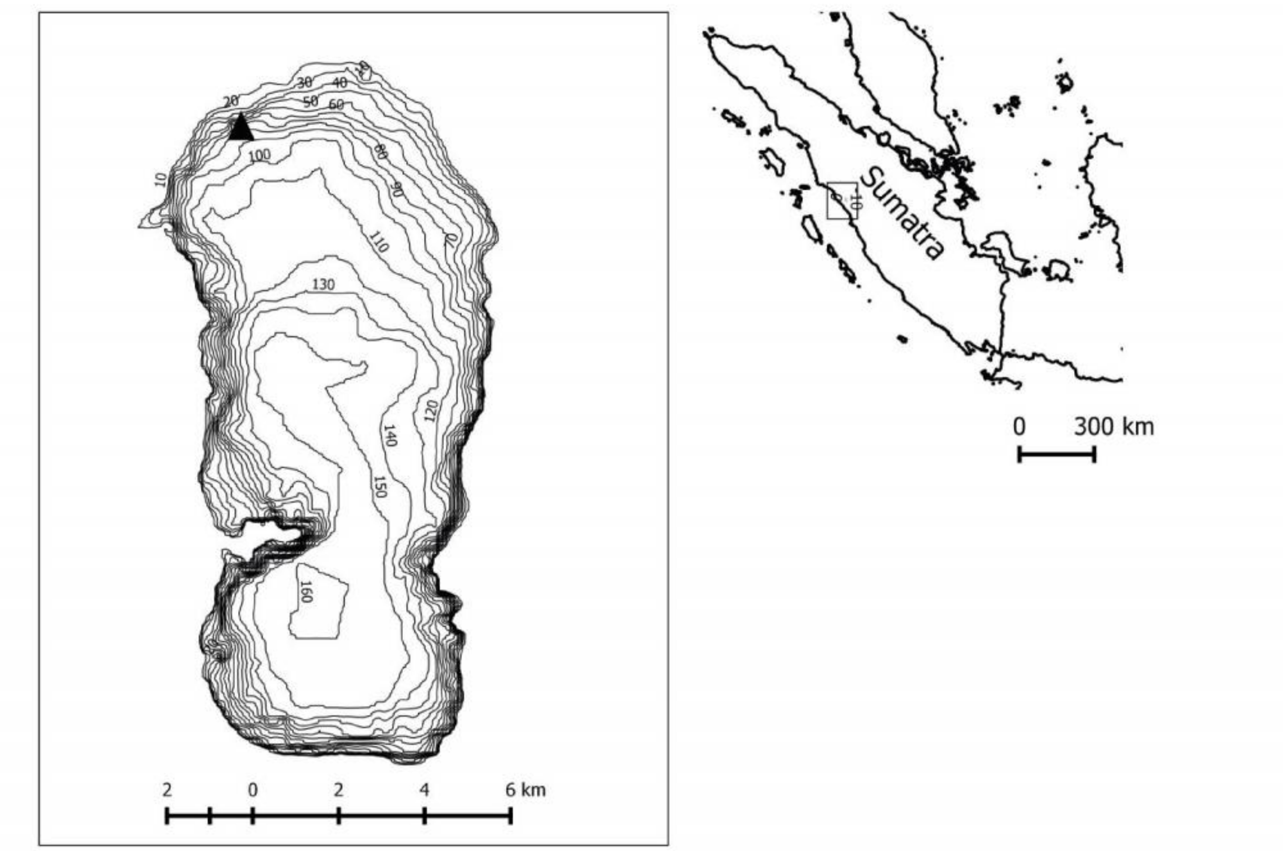

Figure 1. Lake Maninjau in West Sumatra. Location of the monitoring station (E-MOST) is indicated by black triangle. Contour lines indicate bathymetric depth of the lake.

Gambar 1. Danau Maninjau di Sumatra Barat. Lokasi stasiun pengamatan (E-MOST) ditunjukkan oleh segitiga hitam. Garis kontur menunjukkan kedalaman batimetri danau.

Results

Observed water column temperature

Over 31000 temperature profiles data observed by the E-MOST (Table 1) were analysed. Epilimnion temperature of the lake ranged slightly, 26.5 - 28.8 °C, with a mean of 27.2 °C. Hypolimnion temperature was relatively stable at 26.3 °C, only swung 0.2 °C. Hence, the maximum temperature difference between surface and bottom layers was only 2.6 °C. Most water temperature profiles, however, did not show a smooth shape of thermal stratification (e.g. Figure 3).

Data quality and algorithm output

The observed data were relatively in good agreement with theoretical temperature profiles generated by the algorithm (Table 1 and Figure 4). More than 96% of the data followed the shape of that typical profile, although not that smooth (e.g. Figure 3). Root mean square error of the deviated data was calculated at 0.08 ^{o}C. The estimated thermocline depth (Z_{mix}) was normally distributed between 2.2 to 30.0 m (Figure 5). The most common thermocline was identified around 15 m. The steepness of the transition depth (S) was skewed to small values (<5), peaked at 2.08.

Algorithm performance

Owing to the massive resolution of the data to handle (i.e. 31958 temperature profiles), the dataset was separated into a 4-days timeseries profile, from 25-29 April 2018 (n = 589), to highlight the performance of the algorithm. It is clear that the algorithm successfully refined the raw dataset (Figure 6). The raw data (Figure 6A) were smoothed by the algorithm into better referenced based temperature profiles (Figure 6B). From this dataset, it was calculated that there was a relatively constant deviation occurred ~ 6 m depth (RMSE ~0.35 ^{o}C, Figure 7). Some fair errors (RMSE ~0.2 ^{o}C) were also distributed around the estimated thermocline and the interfaces of metalimnion with the adjacent layers (see Figure 3).

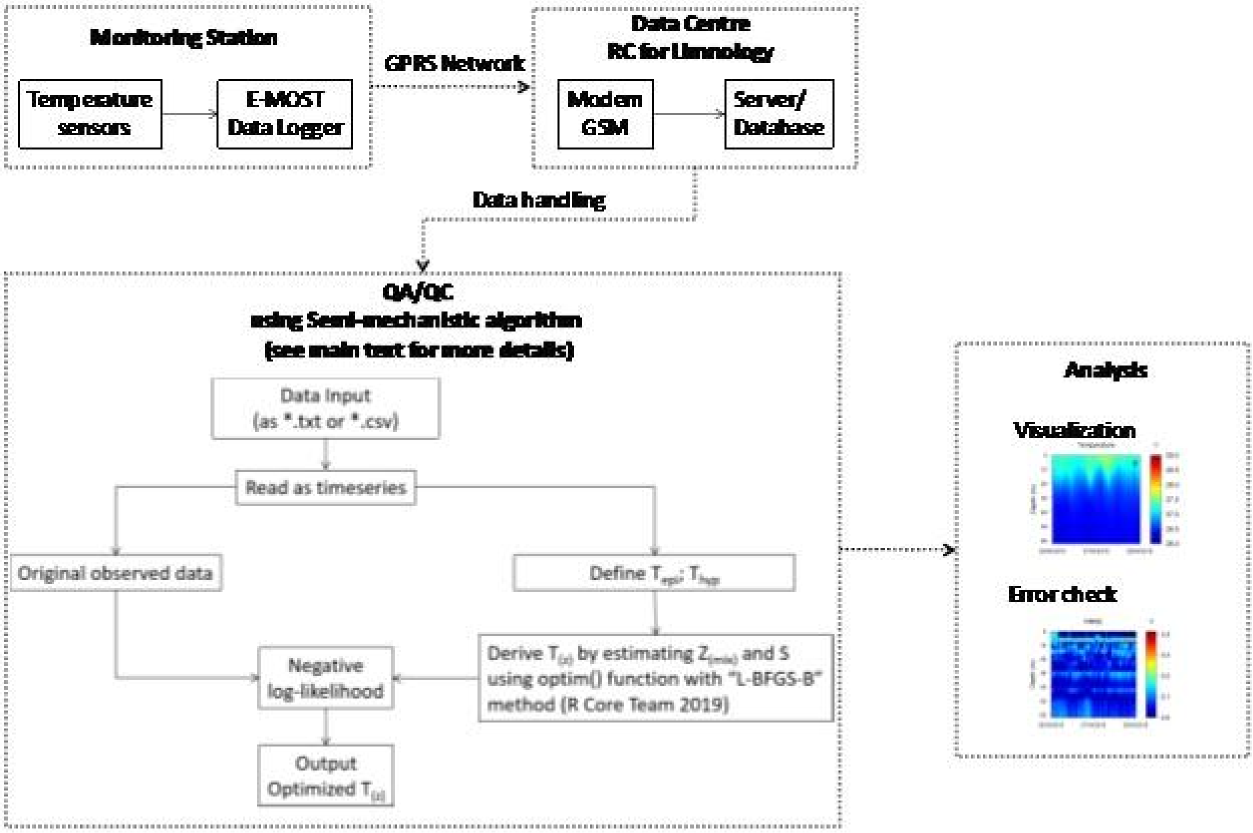

Figure 2. The schematic diagram of data flow from monitoring sensors to desk analysis. QA/QC is introduced to better shape data quality.

Gambar 2. Diagram skematik alur penerimaan data dari sensor pengamatan hingga analisis lebih lanjut. QA/QC diterapkan dalam memperbaiki kualitas data.

Table 1. Observed water column temperature and algorithm output. Values in the parameters are the mean of all sensor observations and numbers in parentheses indicate the range of values. See main text for details.

Tabel 1. Suhu kolom air hasil pengamatan dan luaran algoritma. Nilai parameter merupakan rerata dari seluruh observasi sensor dan angka dalam kurung adalah rentang variasinya. Lihat teks utama untuk keterangan lebih lanjut.

Parameter	Value	Units
T_{epi}	27.2 (26.5 - 28.8)	^{o}C
T_{hyp}	26.3 (26.2 - 26.4)	^{o}C
Z_{mix}	15.2 (2.2 - 30.0)	m
S	2.4 (<0.001 - 50.0)	-
Numbers of temperature profile	31958	-
RMSE	0.08	^{o}C
R^2	0.962	-

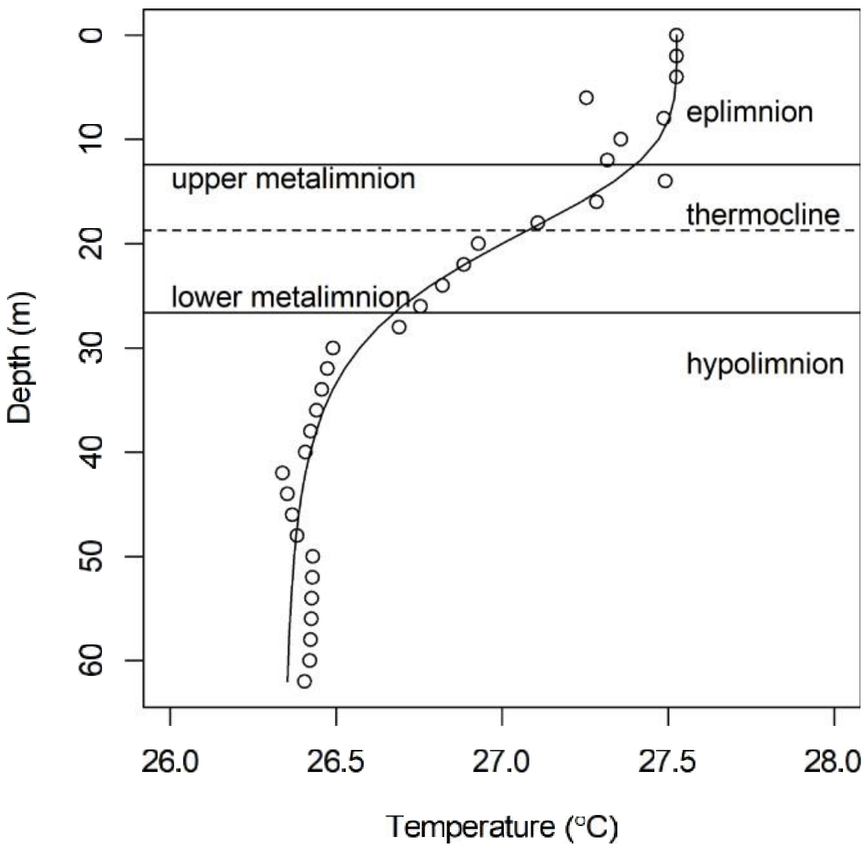

Figure 3. Water temperature profile of Lake Maninjau with its stratification layers, snapshot at 00:45:00 on 1 January 2018. Open circles are the observed temperature by the sensors. Curved solid line is the temperature profile generated by the semi-mechanistic algorithm.

Gambar 3. Profil suhu air Danau Maninjau dan stratifikasinya, diamati pada pukul 00:45:00 tanggal 1 Januari 2018. Lingkaran-lingkaran terbuka merupakan suhu observasi oleh sensor. Garis kurva adalah profil suhu yang dihasilkan oleh algoritma semi-mekanistik.

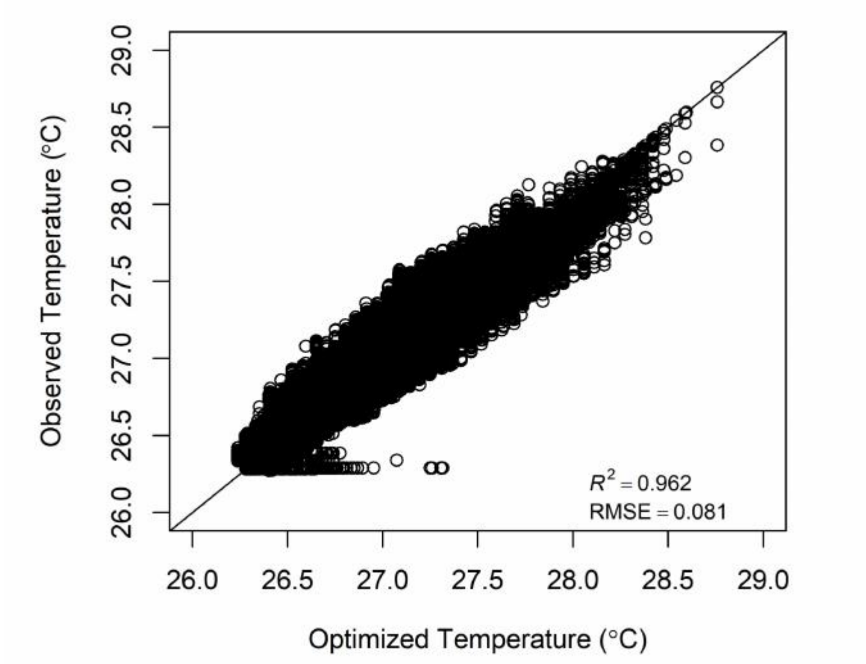

Figure 4. Temperature in the water column of Lake Maninjau observed from 1 January 2018 to 28 May 2018 vs. its optimized temperature generated by the semi-mechanistic algorithm. Solid diagonal line indicates 1:1 relationship.

Gambar 4. Suhu kolom air Danau Maninjau pada tanggal 1 Januari 2018 hingga 28 Mei 2018 sebelum dan sesudah optimasi dengan algoritma semi-mekanistik. Garis diagonal menunjukkan hubungan 1:1.

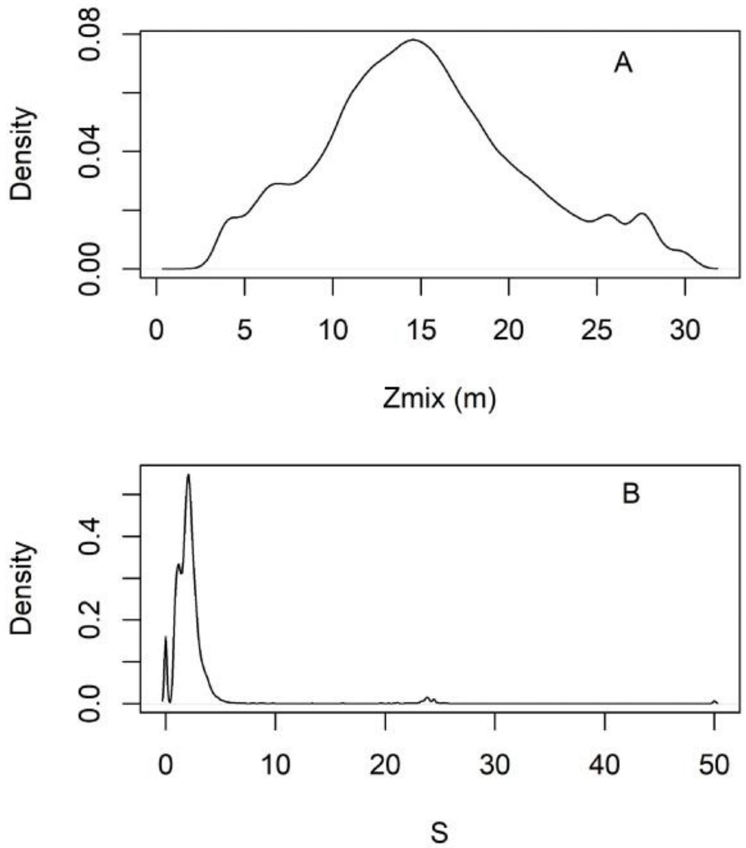

Figure 5. Density plot of algorithm outputs. (A) Estimated thermocline depth (Z_{mix}); (B) The steepness of transition depth between epilimnion and hypolimnion (S).

Gambar 5. Diagram densitas data hasil estimasi algoritma. (A) Kedalaman termoklin (Z_{mix}); (B). Derajat kecuraman antara kedalaman epilimnion dan hipolimnion.

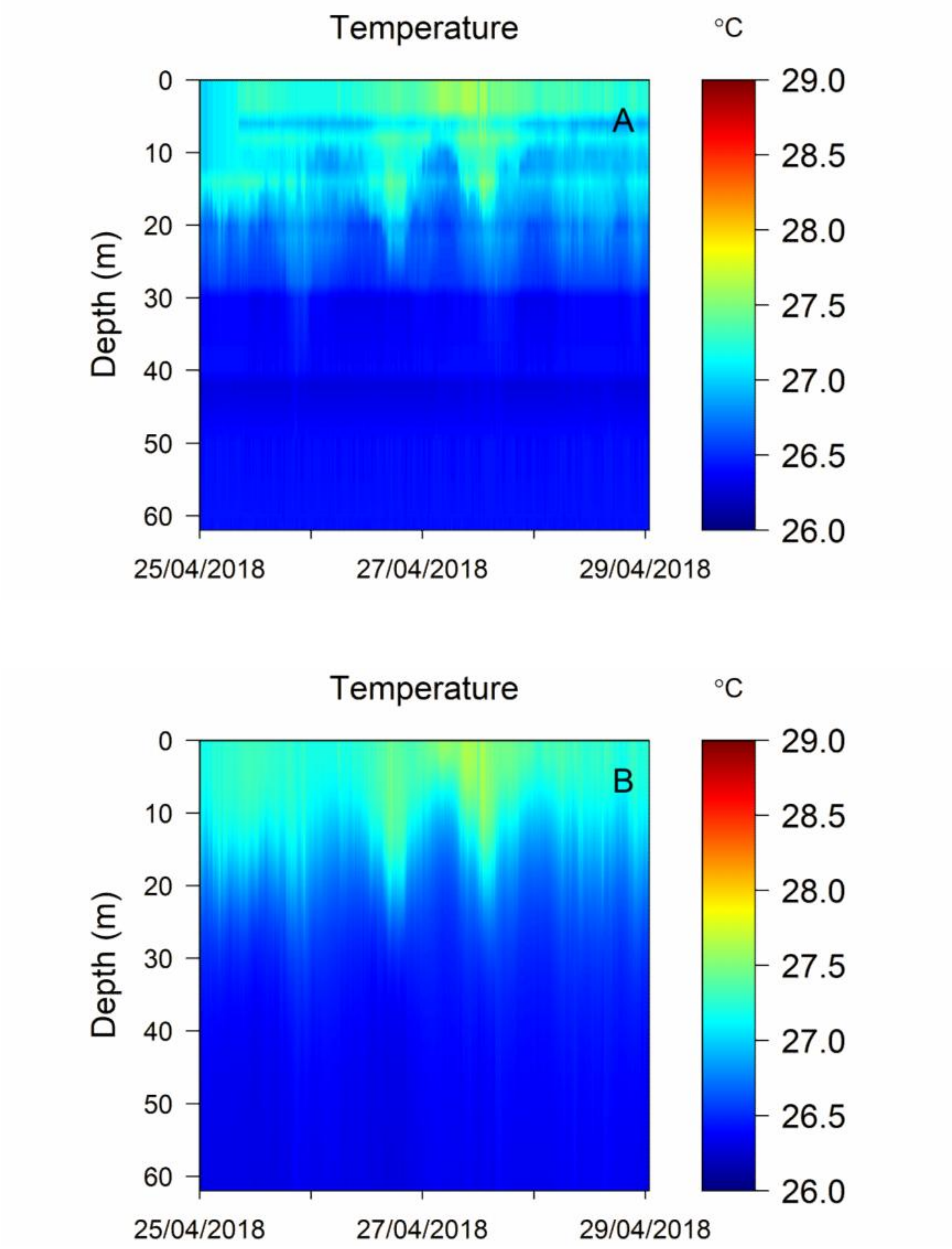

Figure 6. A time-series water temperature profile of Lake Maninjau, snapshot from 25 - 29 April 2018. A. Extracted from raw data, B. Displayed after semi-mechanistic algorithm optimization.

Gambar 6. Profil suhu air Danau Maninjau yang diamati pada tanggal 25 hingga 29 April 2018. A. Diekstrak langsung dari data mentah. B. Digambarkan setelah optimasi dengan algoritma semi-mekanistik.

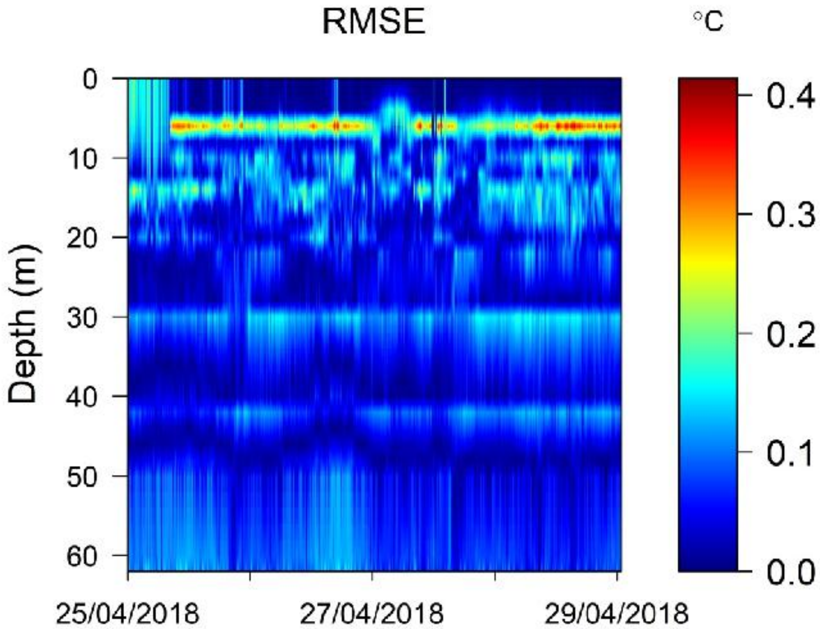

Figure 7. A snapshot of possible errors in the sensors calculated as the deviation of residuals (RMSE) between observed vs. optimized data.
Gambar 7. Kemungkinan eror yang dihitung berdasarkan standar deviasi (RMSE) antara pengamatan dan optimasi data.

Discussion

Environmental data streamed from monitoring sensors may be in a poor qualities condition for many reasons, e.g. system malfunction, drift in reading, errors in transmission (Ganesan et al., 2004). Several procedures of QA/QC could be proceeded to prevent and fix errors and to produce high-quality data (Campbel et al., 2013). These include physical system maintenance, both sensors and network, and post evaluation of the produced data. Under this rule, proceeded on a regular basis, scientists and all other end-users will highly trust the data, particularly in synthesizing crucial information.

High-frequency data and the implemented algorithm in this study have demonstrated the importance of QA/QC procedure. As shown in figure 3, temperature profile depiction from a set of temperature sensors in a thermistor chain of a monitoring station will not as smooth as that from a CTD's cast data. Each sensor in that chain behaves independently and may also drift accordingly producing variation of errors along the line. While a single temperature sensor in the CTD is not likely to produce variations of error. However, in deep lakes, including those in the tropics, a typical temperature profile with its thermal stratification can be indicated (Katsev et al., 2010; Santoso et al., 2018; Wetzel, 2001), although it may not be smooth (Figure 3).

It has been shown that faulty data could be generated by the sensors (e.g. in Figure 6A). Thus, a simple algorithm could be used as solution to fix the errors (e.g. in Figure 6B). Using such solution, one can identify what errors should be tackled. Taking the dataset of 25-29 April 2018 as an example, it can be argued that sensor at 6 m depth should need better treatment (i.e. recalibrated or replaced) owing to its constantly high error values (Figure 5). Such drift might occur as sensor component deteriorate over time because of age-related processes. Placing replicate sensors (at least three) in one depth could also be an option in maintaining good-quality data, as well as in identifying sensor drift. Sensor calibration, e.g. using CTD's cast data as reference, should also be conducted on a regular basis to prevent bad readings.

Data streams produced by sensor networks can be extremely rapid and voluminous (Porter et al., 2012). In this case study, only 31,958 temperature profiles data of 32 depth points were managed. This equals 1.022.656 data points, and that only came from five months measurement, captured every 10 minutes. Efficient data handling, therefore, is demanding. This includes transport and storage, in which the text format file (*.txt) was used to compress data size. Gap filling and data visualization are the next steps of handling quality control prior to further analysis. Hence, a proper software package friendly to big data (e.g. Matlab, R, Phyton) is required in this case.

High-quality data, particularly temperature, is important in monitoring deep tropical lake system. Due to the small variation magnitude, and happened in sub-daily temporal scales (Katsev et al., 2010; Santoso et al., 2018), there is low tolerance for faulty data. Low-quality data that the sensor produced might result in false information. Again, taking the 25-29 April 2018 dataset as an example, estimating thermocline depth from the raw data would be catastrophic and could end in false information (Figure 8). Assuming that thermocline depth is defined as the greatest rate of density change with depth, one might have estimated that these depths laid at 5 m and might vary hourly, or even less (Figure 8A). That would be impossible as changes in heat balance in lakes would not be that fast, naturally. Thermocline depths estimated by our simple semi-mechanistic algorithm showed a more realistic pattern, in which it varies in a sub-daily clear pattern (Figure 8B). A sudden extreme meteorological event may disturb stratification stability which leads to deepening the thermocline depth (Santoso et al., 2018). After the event, shallow thermocline depth would be impossible to develop shortly, e.g. in Figure 8A. A Deep large lake (e.g. Maninjau) has a large heat capacity due the huge water volume. A small rate of heat flux into the lake will not likely affect its thermal stratification. Diurnal variation of heat flux, together with other meteorological force, however, may drive thermal dynamics in the lake particularly in the upper layer, i.e. the epilimnion (Katsev et al., 2010; Santoso et al., 2018). In stratified lakes, this upper mixed layer is the most dynamic in terms of thermal variation. Hence, drifts from each sensor in the thermistor chain may produce errors and deviate the data from the smooth theoretical stratification pattern (Figure 3). This is why a number of errors were found in this active layer (Figure 7). This example demonstrates that it is necessary to analyse circumstances of what really happened under which our data might lead to false interpretation. Such analysis is very critical in synthesizing lake biogeochemical processes.

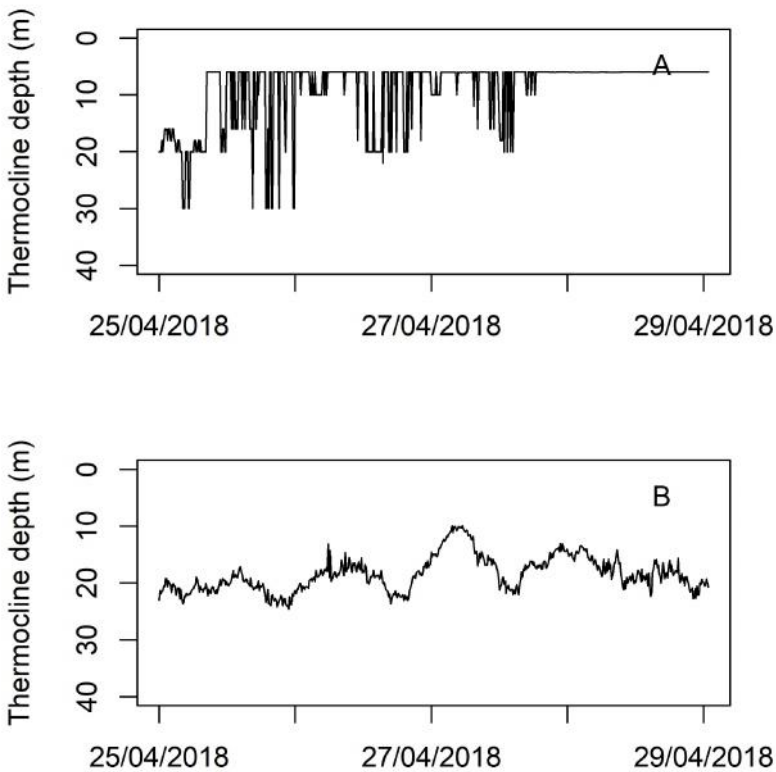

Figure 8. Thermocline depth; A. Estimated from directly from raw sensor data. B. Calculated using semi-mechanistic algorithm optimization as a proxy.

Gambar 8. Kedalaman termoklin. A. Diestimasi dari data mentah sensor. B. Dihitung dengan menggunakan algoritma semi-mekanistik.

Conclusion

This study has demonstrated that QA/QC is necessary for reading high-frequency environmental data logged from monitoring sensors. Through a simple semi-mechanistic algorithm, this study was able to refine lake temperature profiles depicted from the high-frequency data, and also was able to identify a faulty sensor in a particular depth based on its deviation magnitude from the refined profile. This simple solution might assist scientists and all end-users to acquire a better shape of data. As better-quality data would reduce the misleading analysis of processes occurring in lakes, particularly in deep tropical lakes, both, system maintenance and data handling, therefore, are necessary to warrant high-quality information.

Acknowledgement

This study was funded by the State Budget of Indonesian Government (DIPA-APBN) through Indonesian Institute of Sciences. We acknowledge the E-MOST team for initiating the high-frequency monitoring buoy in Lake Maninjau. We would like to thank to UPT Maninjau Team for the field work assistance, and other technical supports, especially in maintaining the monitoring buoy. We also would like to thank the participation of Kabupaten Agam, Ministry for Research and Technology, Ministry for Environment and Forestry, and PT PLN in the development of the monitoring buoy.

References

Adrian, R., O'Reilly, C. M., Zagarese, H., Baines, S. B., Hessen, D. O., Keller, W., & Winder, M. (2009). Lakes as sentinels of climate change. *Limnology and Oceanography*, 54(6 PART 2), 2283–2297. https://doi.org/10.4319/lo.2009.54.6_part_2.2283

Andersen, M. R., de Eyto, E., Dillane, M., Poole, R., & Jennings, E. (2020). 13 years of storms: An analysis of the effects of storms on lake physics on the Atlantic fringe of Europe. *Water*, *12*(2). https://doi.org/10.3390/w12020318

Campbell, J. L., Rustad, L. E., Porter, J. H., Taylor, J. R., Dereszynski, E. W., Shanley, J. B., & Boose, E. R. (2013). Quantity is Nothing without Quality: Automated QA/QC for Streaming Environmental Sensor Data. *BioScience*, *63*(7), 574–585. https://doi.org/10.1525/bio.2013.63.7.10

Dugan, H. A., Iestyn Woolway, R., Santoso, A. B., Corman, J. R., Jaimes, A., Nodine, E. R., & Weathers, K. C. (2016). Consequences of gas flux model choice on the interpretation of metabolic balance across 15 lakes. *Inland Waters*, *6*(4), 581–592. https://doi.org/10.5268/IW-6.4.836

Ganesan, D., Cerpa, A., Ye, W., Yu, Y., Zhao, J., & Estrin, D. (2004). Networking issues in wireless sensor networks. *Journal of Parallel and Distributed Computing*, *64*(7),799–814. https://doi.org/10.1016/j.jpdc.2004.03.016

Hamilton, D. P., Carey, C. C., Arvola, L., Arzberger, P., Brewer, C., Cole, J. J., & Brookes, J. D. (2014). A Global lake ecological observatory network (GLEON) for synthesising high-frequency sensor data for validation of deterministic ecological models. *Inland Waters*, *5*(1), 49–56. https://doi.org/10.5268/IW-5.1.566

Hanson, P. C., Weathers, K. C., & Kratz, T. K. (2016). Networked lake science: how the Global Lake Ecological Observatory Network (GLEON) works to understand, predict, and communicate lake ecosystem response to global change. *Inland Waters*, *6*,543–554. https://doi.org/10.5268/IW-6.4.904

Hipsey, M. R., Hamilton, D. P., Hanson, P. C., Carey, C. C., Coletti, J. Z., Read, J. S., … Brookes, J. D. (2015). Predicting the resilience and recovery of aquatic systems : a framework for model evolution within environmental observatories. *Water Resources Research*, 1–21. https://doi.org/10.1002/2015WR017175

Hipsey, M. R., Bruce, L. C., Boon, C., Busch, B., Carey, C. C., Hamilton, D. P., … Winslow, L. A. (2019). A General Lake Model (GLM 3.0) for linking with high-frequency sensor data from the Global Lake Ecological Observatory Network (GLEON). *Geoscientific Model Development*, *12*(1), 473–523. https://doi.org/10.5194/gmd-12-473-2019

Katsev, S., Crowe, S. A., Mucci, A., Sundby, B., Nomosatryo, S., Haffner, G. D., & Fowle, D. A. (2010). Mixing and its effects on biogeochemistry in the persistently stratified , deep , tropical Lake Matano, Indonesia. *Limnology and Oceanography*, *55*(2), 763–776.

Mackenzie-Grieve, J. L., & Post, J. R. (2006). Projected impacts of climate warming on production of lake trout (*Salvelinus namaycush*) in southern Yukon lakes. *Canadian Journal of Fisheries and Aquatic Sciences*, *63*(4), 788–797. https://doi.org/10.1139/f05-257

Minns, C. K., Shuter, B. J., & Bay, J. (2013). A semi-mechanistic seasonal temperature-profile model (STM) for the period of stratification in dimictic lakes. *Canadian Journal of Fisheries and Aquatic Sciences*, *181*, 169–181.

Porter, J. H., Hanson, P. C., & Lin, C. C. (2012). Staying afloat in the sensor data deluge. *Trends in Ecology and Evolution*, *27*(2), 121–129. https://doi.org/10.1016/j.tree.2011.11.009

R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from https:// www.R-project.org/.

Read, J. S., Hamilton, D. P., Jones, I. D., Muraoka, K., Winslow, L. a, Kroiss, R., … Gaiser, E. (2011). Derivation of lake mixing and stratification indices from high-resolution lake buoy data. *Environmental Modelling & Software*, *26*(11), 1325–1336. http://dx.doi.org/10.1016/j.envsoft.2011.05.006

Santoso, A. B., Triwisesa, E., Fakhrudin, M., Harsono, E., & Rustini, H. A. (2018). What do we know about Indonesian tropical lakes? Insights from high frequency measurement. *IOP Conference Series: Earth and Environmental Science*, *118*(1). https://doi.org/10.1088/1755-1315/118/1/012024

Wetzel, R. G. 2001. *Limnology: Lake and river ecosystems*. New York: Elsevier.

Winslow, L., Read, J., Woolway, R., Brentrup, J., Leach, T., & Zwart, J. (2015). rLakeAnalyzer: Package for the Analysis of Lake Physics. R package version 1.7.3. Retrieved from http://cran.r-project.org/ package=rLakeAnalyzer