Turkish Journal of Water Science & Management
ISSN: 2536 474X / e-ISSN: 2564-7334

Volume: 6 Issue: 2 Year: 2022

Research Article

The Effect of Data Granularity on Temperature Gradient Modeling in Michigan's Streams

Veri Taneselliğinin Michigan Akarsularının Sıcaklık Gradyan Modellemesi Üzerindeki Etkisi

Halil I. Dertli[1], Daniel B. Hayes[2], Troy G. Zorn[3]
[1]*Republic of Türkiye Ministry of Agriculture and Forestry, the General Directorate of Water Management, Beştepe, Söğütözü St. No:14, Yenimahalle, Ankara, TURKIYE 06560*
dertliha@msu.edu (https://orcid.org/0000-0003-2311-8741)
[2]*Michigan State University, Department of Fisheries and Wildlife, Natural Resources Building, 480 Wilson Rd. East Lansing, Michigan, USA 48824*
hayesdan@msu.edu (https://orcid.org/0000-0002-8132-4749)
[3]*Michigan Department of Natural Resources, Marquette Fisheries Research Station, 484 Cherry Creek Road, Marquette, Michigan, USA 49855*
zornt@michigan.gov (https://orcid.org/0000-0001-7552-6398)
Received Date: 14.03.2022, Accepted Date: 07.06.2022
DOI: 10.31807/tjwsm.1084423

Abstract

Stream temperature is a critical characteristic for aquatic ecosystems. Therefore, it is crucial to understand the factors that take place in thermodynamic processes in these ecosystems. Regression models are useful tools that help us comprehend and explain the drivers of these thermal processes since they can be used for quantifying the magnitude and the type of the relationship between the independent variables (e.g., air temperature, discharge) and the response variable (e.g., stream temperature). However, selection of data granularity of data may often be a key decision for modelers. Although granularity of data is selected based on the ecological relevance of data to the question of interest in many cases, it may arbitrarily be selected by the researchers in many other cases. However, data granularity can substantially influence model coefficients, can affect the model predictions, and influence evaluation of model fitness and interpretation of model outputs. In this article, we adopted regression models and applied different data granularity scenarios to investigate the consequences of data granularity selection in modeling approaches. Our findings showed that using different data granularities resulted in considerable changes in regression coefficients in the models. Our results also revealed that overall model fitness increased with coarser-scale data granularity and model selection was influenced by the type of data granularity. This study might be helpful for modelers and environmental managers since it highlights the significance of selection of data granularity and proposes a different point of view in model design, evaluation and application from the perspective of data selection.

Keywords: *stream temperature, linear regression models, data granularity, data aggregation, temporal scale*

Öz

Akarsu sıcaklıkları sucul ekosistemlerde kritik öneme sahiptir. Dolayısıyla, akarsulardaki termodinamik süreçleri etkileyen faktörleri kavramak önem arz etmektedir. Regresyon modelleri bağımsız (örn. havanın sıcaklığı, akı) ve bağımlı (örn. akarsu su sıcaklığı) değişkenlerin birbirleriyle olan nicel ve nitel ilişkisini açıklayabildiğinden, bu ısıl süreçleri etkileyen faktörleri kavramamıza ve açıklamamamıza yardımcı olan kullanışlı araçlardır. Ancak bu modellerde kullanılan verilerin taneselliğinin ya da agregasyonun seçimi modellemeciler için zorlayıcı olabilmektedir. Çoğu durumlarda kullanılacak verinin taneselliği ekolojik uygunluğa bağlı olarak seçilse de diğer birçok durumda keyfi olarak seçilebilmektedir. Ancak veri taneselliği seçimi, model değişkenlerinin katsayılarını, model tahminlerini, model tahminlerinin değerlendirilmesini ve model sonuçlarının yorumlanmasını önemli ölçüde etkileyebilmektedir. Bu makalede, veri taneselliği seçiminin etkilerini araştırmak amacıyla regresyon modelleri farklı veri taneselliği senaryolarıyla uygulandı. Bulgular, veri taneselliği seçiminin regresyon değişken katsayılarını önemli düzeyde etkilediğini gösterdi. Ayrıca bulgular veri taneselliğindeki artışın ortalama model tahmin gücünü artırdığını ve veri taneselliğinin model seçimlerinde etkili olduğunu ortaya çıkardı. Bu çalışma veri taneselliği seçiminin önemini vurgulaması, model tasarımı, değerlendirilmesi ve uygulanması konularında farklı bir bakış açısı sunması sebebiyle, modellemecilere ve yöneticilere yararlı olabilir.

Anahtar sözcükler: *akarsu sıcaklığı, doğrusal regresyon modelleri, veri taneselliği, veri kümelenmesi, zamansal ölçek*

Introduction

Stream temperature plays a key role in the physical, chemical, and biological dynamics in freshwater ecosystems. Therefore, it is often considered as one of critical parameters in evaluation of water quality and ecosystem functioning in the literature (Neumann et. al., 2003; Ducharne, 2008; Ficklin et. al., 2013; Guo et. al., 2019; Hamid et. al., 2020). As water temperature influences the survival, reproduction and distribution of species from different taxa (e.g., primary/secondary producers, aquatic invertebrates, fish and other aquatic vertebrates), it is crucial to understand the physical determinants of water temperature in these ecosystems (Iversen, 1971; Jackson et. al., 2007; Zorn et. al., 2004; Nuhfer et. al., 2017).

In literature, various environmental parameters are used to explain the driving factors of stream temperatures. Du et. al. (2020), for example, propose that both meteorological (e.g., air temperature) and hydrological (e.g., precipitation) processes affect stream temperatures. In other studies, these meteorological and hydrological processes are diversified into different sub-factors. For example, Cheng and Wiley (2016) describe the radiative processes such as shortwave and longwave solar radiation as explanatory meteorological factors in thermal dynamics of streams (Figure 1). Hydrological characteristics such as water depth, surface area, runoff, and

groundwater contribution/withdrawal are also included as the key processes that determine the thermodynamics in a stream (Zorn et. al., 2008; Cheng, & Wiley, 2016; Du et. al., 2020; Andrews, 2019; Dertli, 2021). As the stream ecosystems are open systems, all these processes interact with each other, which makes understanding individual roles of these physical processes in stream thermodynamics hard to comprehend for researchers. At this point, statistical models help researchers explain these roles in these complex natural systems.

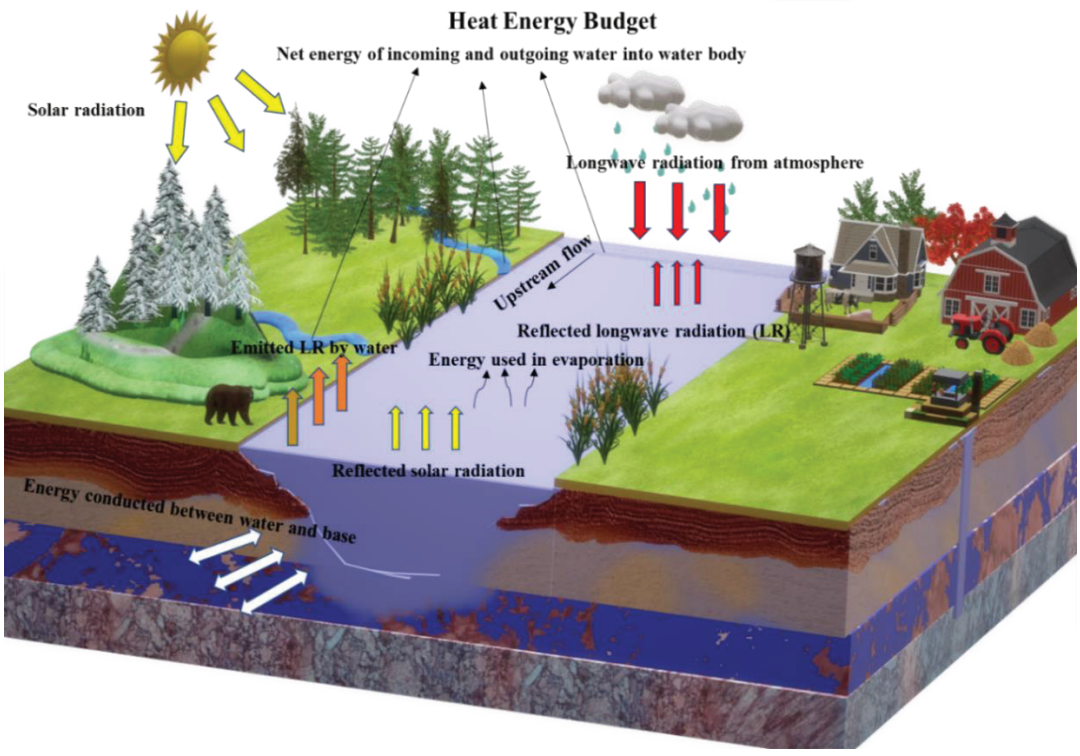
Statistical models are frequently used to understand the thermal dynamics in freshwater ecosystems. Regression models, for example, are able to quantify the influence of each parameter in the model on the response variable (Bender, 2009). Therefore, these models are very useful to evaluate the potential effects of different stress factors (e.g., climate change, groundwater withdrawal) on these valuable systems as they can make predictions on the trends of thermal dynamics under alternating environmental conditions (Mantua et. al., 2010; Andrews, 2019). Once successfully designed for a certain group of streams (e.g., cold streams), statistical models can reduce the need for extensive data collection, which can reduce the financial resources, time and labor that are spent in data collection procedures (Dertli, 2021). In addition, model predictions can be useful in making future projections on the population dynamics of various aquatic organisms such as fish (Chang et. al., 2018; Nuhfer et. al, 2017), and play critical roles in decision-making processes on environmental issues.

Although statistical models are useful tools for understanding the nature of thermodynamics in streams, the explanatory power of these models may depend on the structure of input data (Akossou & Palm, 2013). The type of time aggregation, or data granularity –defined as a new term in environmental studies by Dertli (2021)– is one of the important structural features of the data, since it directly influences the number of data points (e.g., sample size) and the collinearity between the model parameters (Stefan & Preud'homme, 1993; Pilgrim et. al., 1998). Because it can change the outputs of regression models, data granularity is often important. In literature, selection of data granularity is generally based on the ecological relevance of the selected data granularity to the research question of interest, and researchers often provide strong reasoning for data granularity selected in their studies. For example, Chen et. al. (1998) adopt hourly data granularity to simulate stream temperatures based on the shading dynamics of topography and vegetation throughout the day. In another study, Zorn et. al. (2004) focus on July mean temperatures as a reference temperature for Michigan streams because of it indexed

conditions important to fish growth, survival, and abundance. However, in many other cases, researchers arbitrarily select the type of data granularity used in their models, even though arbitrary selection may cause misevaluations of model predictions and biases in model selection processes (Dertli, 2021).

Figure 1

Environmental Processes That are İnvolved in Stream Thermodynamics (Dertli, 2021)



So far, different studies adopt different approaches on the issue of data granularity selection, develop different perspectives and reveal various consequences of these selections (Stefan, & Preud'homme, 1993; Pilgrim et. al., 1998; Webb et. al., 2003). However, there are still only a few studies that focus on this issue, considering the substantial effects of data granularity selection on evaluation, selection, and interpretation of linear regression models. Therefore, in this paper, we focus on the response of linear regression models designed by Andrews (2019) to simulate effects of streamflow on temperature gradient (i.e.,

change in water temperature between upstream and downstream locations) to changes in granularity of the data used in models. Our study objectives are:

1. To evaluate and interpret the response of model coefficients to different data granularity scenarios,
2. To evaluate the fitness of the regression models under different data granularity scenarios,
3. To evaluate influence of data granularity selection on the selection of the most parsimonious (i.e., high model fitness with low model complexity) model.

Since we adopt different approaches and evaluate the models based on different characteristics (e.g., model fitness and parsimony) to observe the response of regression models to different granularity scenarios, this paper can give researchers a broad perspective on possible consequences of their data granularity selection.

Methods

Study Site, Data Collection and Data Revision

The streams were selected by Andrews (2019) for data collection throughout State of Michigan. Andrews (2019) collected data from 21 streams with various periods (e.g., between July and November) in 2015 and 2016 (Table 1). He collected water temperature and water pressure data by setting HOBO® U20 Water Level Loggers at both upstream and downstream data collection points. These data were collected in 15-minute intervals and averaged into hourly interval. Water pressure data were used to calculate upstream and downstream discharge after obtaining stream width and stream depth estimations for both upstream and downstream data collection stations. Stream velocity data were also collected for both stations by using SonTek® Flowtracker. Methods for discharge calculations are explained in the study of Andrews (2019) in details. In addition, Andrews (2019) collected air temperature and barometric pressure data from paired streams that were located in close range by using Monarch® Track-It data loggers with 15-minute time intervals. These data were also averaged into hourly interval.

In addition to air temperature, water temperature and discharge values, Andrews (2019) calculated other environmental variables, such as altitude angle, to

use them as model parameters. Calculations for these environmental variables are explained in the study of Andrews (2019) in details. We obtained data for all environmental variables in hourly time interval from Andrews's (2019) study to use in our study.

We revised the data by detecting and eliminating outliers, also by trimming the data within June-October period for both 2015 and 2016. We selected this period for our study since it covers summer season, which is important for fish abundance (Zorn et. al. 2004). Another reason was that this period was the longest range of data that is found commonly for all streams. Since most of the stream's data started from late July in 2015, we only used 2016 data in this study (Dertli 2021). In addition, we used the data from 16 out of 21 streams in this study to avoid gaps in data that were detected in some streams (Figure 2).

After data revisions, we aggregated the hourly data by averaging the observations into 2-hours, 6-hours, 12-hours, daily and weekly time intervals. In the end, we obtained 1-hour (hourly), 2-hours, 6-hours, 12-hours, 24-hours (daily) and 168-hours (weekly) data granularity scenarios.

Hierarchical Model Development and Model Simulation

Andrews (2019) designed 11 linear regression models to obtain temperature gradient predictions (the difference between downstream and upstream water temperatures). He adopted hierarchical model development, in which models were formed starting from the least complex (i.e., Model 1) to the most complex (i.e., Model 10). At each step, a new parameter was included in the model, or an existing model parameter was replaced with another model parameter (Table 2). Model 11, however, was adopted from a physical model that was proposed by Magnusson et. al. (2012). In our study, these models were simulated for each stream and each data granularity scenario.

Model Fitness and Selection

We used adjusted correlation coefficient ($R^2_{adj.}$) to evaluate the amount of fit between the trends of observed and predicted temperature gradient (ΔT). Adjusted correlation coefficient was calculated based on the equation:

$$R^2_{adj.} = 1 - \frac{SSE/(n-p)}{SST/(n-1)},$$

where *n* stands for the number of observations, *p* stands for the number of parameters, *SSE* and *STT* stand for sum of squared residuals error and total sum of squares, respectively. To find the most parsimonious model under given conditions, we used model weight (ω) of models for each stream. To find model weights, we obtained Akaike's Information Criterion (AIC) values based on Akaike (1973). AIC values were obtained by using the equation:

$$AIC = 2k - 2\,ln\,ln\left(L(data)\right);\ \text{and}\ L(data) = -(\frac{n}{2}) \cdot log_e(SSE),$$

where *L* stands for the likelihood, *k* stands for the unknown parameters and n stands for the sample size (Seber, & Wild, 1989). We used AIC values to obtain model weight as shown in the equation:

$$\omega_i = \frac{exp(-\frac{\Delta i}{2})}{\sum_m^M exp(-\frac{\Delta i}{2})},$$

where *M* is the total number of models, *m* is the model number, and Δi is the difference between AIC values of model *i* and the AIC value of the best-fitting model (Andrews, 2019). By using model weight, we compared the explanatory power of models and their model complexity based on the law of parsimony.

Table 1

List of the Streams That Are Used in This Study (Andrews, 2019)

Stream	Abbr.	Region	Upstream Latitude	Upstream Longitude	Downstream Latitude	Downstream Longitude
Pokagon Creek	PK	SLP	41.89517	-86.162632	41.915803	-86.175679
Pigeon River	PG	SLP	42.932887	-86.081828	42.91636	-86.146075
Nottawa Creek	NTW	SLP	42.192564	-85.060415	42.195998	-85.104618
Tobacco River	TB	SLP	43.909194	-84.697312	43.929905	-84.666327
Hasler Creek	HS	SLP	43.042332	-83.423206	43.083594	-83.442947
Prairie River	PR	SLP	41.801832	-85.116614	41.832568	-85.165065
Swan Creek	SW	SLP	41.90477	-85.297885	41.921249	-85.312047
Cedar Creek	CC	NLP	44.375846	-85.972647	44.369588	-85.999598
Cedar River	CR	NLP	44.956875	-85.132748	44.968664	-85.138993
Black River	BL	NLP	45.070651	-84.283728	45.089439	-84.284929
Butterfield Creek	BF	NLP	44.273249	-85.094087	44.256377	-85.03362
Morgan Creek	MG	UP	46.519698	-87.504502	46.521351	-87.494782
Spring Creek	SP	UP	46.512909	-90.156133	46.513418	-90.177011
Carp River	CP	UP	46.509131	-87.418924	46.510534	-87.388497
Escanaba River	ESC	UP	46.420206	-87.797962	46.398398	-87.770883
Squaw Creek	SQ	UP	46.057035	-87.18974	45.985396	-87.140559

Figure 1

Locations of Streams That Are Used in This Study (Dertli, 2021)

Table 2

List of Multiple Linear Regression Models (Magnusson et. al., 2012; Andrews, 2019)

Model 1	$\Delta T = \beta_0 + \beta_1(T_a - T_w)$
Model 2	$\Delta T = \beta_0 + \beta_1(T_a - T_w) + \beta_2\left(\frac{Q_{up}}{Q_{down}}\right)$
Model 3	$\Delta T = \beta_0 + \beta_1(T_a - T_w) + \beta_3\left(Q_{up}\right) + \beta_4(\Delta T_{flow})$
Model 4	$\Delta T = \beta_0 + \beta_1(T_a - T_w) + \beta_3\left(Q_{up}\right) + \beta_4(\Delta T_{flow}) + \beta_5\left(Q_{down} - Q_{up}\right)$
Model 5	$\Delta T = \beta_0 + \beta_1(T_a - T_w) + \beta_3\left(Q_{up}\right) + \beta_4(\Delta T_{flow}) + \beta_5\left(Q_{down} - Q_{up}\right) + \beta_6(S)$
Model 6	$\Delta T = \beta_0 + \beta_1(T_a - T_w) + \beta_3\left(Q_{up}\right) + \beta_4(\Delta T_{flow}) + \beta_5\left(Q_{down} - Q_{up}\right) + \beta_7(\alpha)$
Model 7	$\Delta T = \beta_0 + \beta_1(T_a - T_w) + \beta_3\left(Q_{up}\right) + \beta_4(\Delta T_{flow}) + \beta_5\left(Q_{down} - Q_{up}\right) + \beta_6(S) + \beta_7(\alpha)$
Model 8	$\Delta T = \beta_0 + \beta_1(T_a - T_w) + \beta_3(Q_{up}) + \beta_5(Q_{down} - Q_{up}) + \beta_6(S) + \beta_8(\Delta T_{up}) + \beta_9(\Delta T_{base}) + \beta_{10}(\Delta T_{over})$
Model 9	$\Delta T = \beta_0 + \beta_1(T_a - T_w) + \beta_3(Q_{up}) + \beta_5(Q_{down} - Q_{up}) + \beta_7(\alpha) + \beta_8(\Delta T_{up}) + \beta_9(\Delta T_{base}) + \beta_{10}(\Delta T_{over})$
Model 10	$\Delta T = \beta_0 + \beta_1(T_a - T_w) + \beta_3(Q_{up}) + \beta_5(Q_{down} - Q_{up}) + \beta_6(S) + \beta_7(\alpha) + \beta_8(\Delta T_{up}) + \beta_9(\Delta T_{base}) + \beta_{10}(\Delta T_{over})$
Model 11	$\Delta T = \beta_0 + \beta_1(T_a - T_w) + \beta_9(\Delta T_{base}) + \beta_{10}(\Delta T_{over}) + \beta_{11}(\frac{1}{Q_{up}} * \left((T_a + 273.16)^4 + (T_w + 273.16)^4\right)] + \beta_{12}\left[\frac{1}{Q_{up}} * (e^{T_w} - e^{T_a})\right] + \beta_{13}(\frac{1}{Q_{up}} * \alpha)$

Note. ΔT: temperature gradient (℃), T_a: Air Temperature (℃), Tw: Upstream temperature (℃), Q_{up}: Upstream discharge (m^3/s), Q_{down}: Downstream discharge (m^3/s), ΔT_{flow}: Cumulative temperature gradient (℃), S: Day length, α: Altitude angle, ΔT_{up}: Upstream temperature gradient (℃), ΔT_{base}: Baseflow temperature gradient (℃), ΔT_{over}: Overflow temperature gradient (℃).

Results

Regression Coefficients

Regression coefficients were obtained after model simulations for each stream. Only Model 10 coefficient values are shown in Table 3, since previous studies showed that Model 10 had the highest model fit (Andrews, 2019; Dertli, 2021). Regression coefficients of model parameters varied across streams (Table 3). For example, the air temperature-upstream temperature gradient ($T_a - T_w$) parameter coefficient had the value of 28.338 in the Carp River model, but a value of -18.751 in the Prairie River model (Table 3). Likewise, coefficient values of upstream discharge (Q_{up}) ranged between -9.301 (Carp River) and 5.079 (Pokagon Creek).

Table 3

Intercepts and Regression Coefficients in Model 10 for Each Stream by Using Hourly Data Granularity Scenario (Dertli, 2021)

Streams	Intercept (β_0)	T_a-T_w (β_1)	Q_{up} (β_3)	$Q_{down} - Q_{up}$ (β_5)	S (β_6)	α (β_7)	ΔT_{up} (β_8)	ΔT_{base} (β_9)	ΔT_{over} (β_{10})
BL	0.004	-0.380	-0.872	0.002	0.004	-0.037	-0.060	-0.013	-0.026
CR	0.982	-0.665	2.810	-0.154	0.005	-0.011	0.097	-0.029	0.015
CC	-3.800	-0.044	0.102	-0.041	-0.001	0.009	0.004	-0.012	0.001
MG	-0.258	7.516	-0.225	0.155	-0.012	-0.015	-0.231	0.186	-0.027
PK	-2.326	-1.105	-1.405	0.072	-0.014	0.027	0.043	-0.105	0.048
BF	1.690	-0.323	-0.343	0.309	0.053	-0.016	0.020	-0.006	-0.003
CP	0.500	28.338	-9.301	-0.038	0.018	-5.671	0.133	0.056	-0.004
PG	-3.953	1.841	5.079	0.009	-0.020	0.018	-0.038	0.015	0.013
SP	1.284	2.412	1.146	0.147	-0.020	0.002	0.060	-0.094	0.038
ESC	-5.148	1.626	-1.804	0.352	-0.066	0.078	-0.130	0.123	-0.033
NTW	-4.156	-0.436	-0.050	0.245	-0.060	-0.017	0.064	-0.120	0.068
TB	1.092	-1.513	-2.864	0.584	-0.077	-0.018	-0.472	0.370	-0.340
HS	2.638	0.121	-0.061	-0.037	-0.022	-0.041	0.041	0.004	0.008
PR	-5.764	-18.751	1.618	0.082	-0.038	0.019	0.244	-0.256	0.117
SQ	0.246	-1.359	-2.618	0.265	0.008	-0.019	0.140	-0.108	0.130
SW	-2.335	-2.630	0.366	0.000	0.001	-0.023	-0.088	0.008	-0.060

In addition, regression coefficients were obtained by simulating models under different data granularity scenarios. Regression coefficients of Model 10 for Tobacco River were shown (Table 4) because preliminary results showed that the predictive power of Model 10 had the highest value (R^2_{adj} = 0.778). Using the data with different granularities changed the regression coefficient values and signs for the same stream (Table 4). For example, regression coefficient of air temperature-upstream temperature gradient ($T_a - T_w$) had the value of 0.023 under 1-hour data

granularity scenario, whereas the same coefficient had the value of -0.023 under 24-hour data granularity scenario. As another example, coefficient of discharge gradient ($Q_{down} - Q_{up}$) variable was -0.579 under 1-hour scenario, while the same coefficient had the value of 0.366 under 24-hour scenario.

Table 4

Regression Coefficient Values of Variables in Model 10 for Tobacco River (Dertli, 2021)

Data Granularity	β_0	β_1	β_3	β_5	β_6	β_7	β_8	β_9	β_{10}
1-hour	0,627	0,023	-1,166	-0,579	-0,006	0,019	-0,223	0,138	-0,144
2-hour	0,629	0,023	-1,169	-0,583	-0,006	0,019	-0,224	0,139	-0,145
6-hour	0,868	0,014	-1,43	-0,522	-0,023	0,022	-0,220	0,125	-0,137
12-hour	1,337	0,001	-2,045	0,242	-0,051	0,023	-0,055	-0,001	-0,031
24-hour	1,092	-0,023	-2,630	0,366	0,000	0,001	-0,088	0,008	-0,060
168-hour	0,780	-0,029	-1,383	-0,436	-0,056	0,000	-0,071	-0,016	-0,067

Model Fitness

Mean R^2_{adj} values of all streams for each regression model showed that Model 10 had the highest model prediction power under all data granularity scenarios (Table 5; Figure 3). When mean R^2_{adj} values of regression under all scenarios were averaged, Model 10 had the highest model fit with the average mean R^2_{adj} value of 0.548. In addition, model fitness increased with data granularity in most cases (Figure 3). For example, mean R^2_{adj} of Model 10 increased from 0.418 to 0.842 from 1-hour to 168-hour scenarios. Moreover, average mean R^2_{adj} of all models under 1-hour scenario was 0.255, whereas average mean R^2_{adj} of models under 168-hour scenario was 0.680.

Table 5

Mean Adjusted Correlation ($R^2_{adj.}$) Values of Each Model by Data Granularity across All Streams (Dertli, 2021)

	Data Granularity (h)						
Model	1	2	6	12	24	168	Average
1	0.139	0.142	0.149	0.198	0.315	0.498	0.240
2	0.094	0.098	0.108	0.133	0.202	0.415	0.175
3	0.188	0.209	0.207	0.226	0.311	0.499	0.273
4	0.205	0.209	0.225	0.253	0.340	0.571	0.301
5	0.278	0.284	0.309	0.368	0.502	0.732	0.412
6	0.253	0.257	0.279	0.360	0.485	0.737	0.395
7	0.329	0.336	0.367	0.502	0.515	0.754	0.467
8	0.258	0.375	0.391	0.453	0.591	0.812	0.480
9	0.332	0.336	0.358	0.45	0.587	0.823	0.481
10	0.418	0.423	0.447	0.563	0.598	0.842	0.548
11	0.312	0.320	0.342	0.419	0.536	0.793	0.454
Average	0.255	0.272	0.289	0.357	0.453	0.680	

Model Selection

Results showed that Model 10 had the highest model weight for most of the streams (i.e., 62.5% of all streams) under 1-hour and 2-hours scenarios (Table 6; Figure 4). Model 10 also had the highest percentage under 6-hours (50.00 %), 12-hours (43.75 %) and 168-hours (31.25 %) scenarios. However, Model 11 had the highest percentage (31.25 %) under 24-hours granularity scenario (Table 6). Moreover, as data granularity increased, number of models that had the highest model weight for at least one of the streams increased. For example, there were only 4 models (Model 8, Model 9, Model 10 and Model 11) that appeared to have the highest model weight for at least one stream under 1-hour granularity scenario, yet we observed 6 models (Model 1, Model 5, Model 7, Model 8, Model 10 and Model 11) that had the percentage value greater than zero (Figure 4) under 168-hours granularity scenario.

Figure 2

Mean Adjusted Correlation ($R^2_{adj.}$) Values of Models across Data Granularity Scenarios (Dertli, 2021)

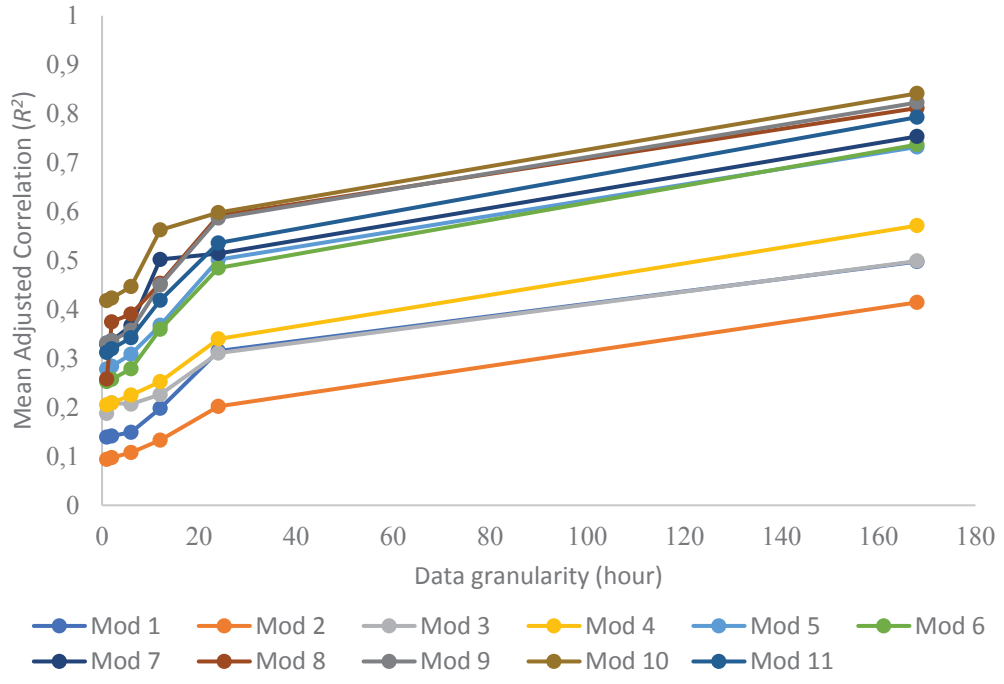


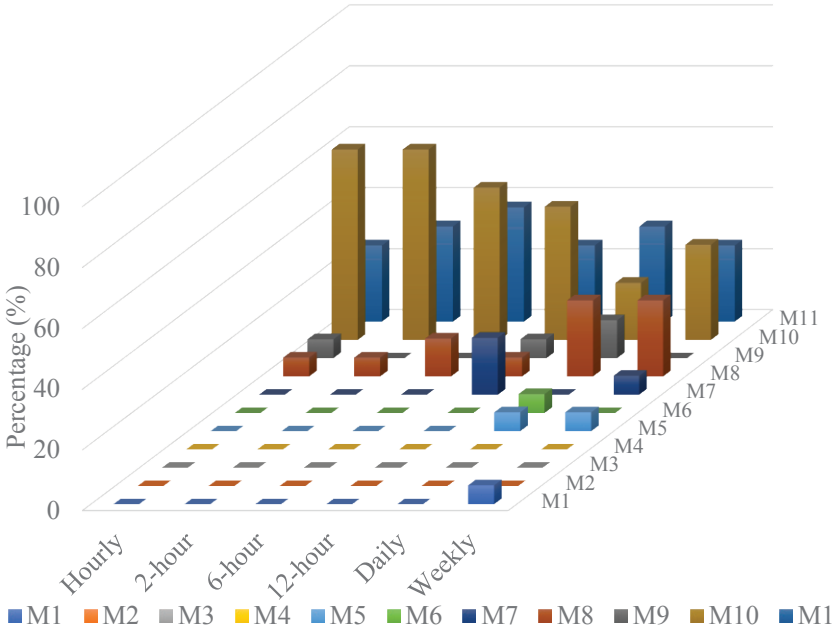
Table 6

Percentage (%) of Streams Where Each Model Had the Highest Model Weight (ω) Across Levels of Data Granularity (Dertli, 2021)

	Models											
Data granularity (hour)	1	2	3	4	5	6	7	8	9	10	11	Total
1	0	0	0	0	0	0	0	6.25	6.25	62.50	25.00	100
2	0	0	0	0	0	0	0	6.25	0	62.50	31.25	100
6	0	0	0	0	0	0	0	12.50	0	50.00	37.50	100
12	0	0	0	0	0	0	18.75	6.25	6.25	43.75	25.00	100
24	0	0	0	0	6.25	6.25	0	25.00	12.50	18.75	31.25	100
168	6.25	0	0	0	6.25	0	6.25	25.00	0	31.25	25.00	100

Figure 3

The Percentage (%) of the Models Having the Highest Model Weight at Least One Stream for Each Data Granularity with June-October 2016 Data (Dertli, 2021)



Discussion and Conclusions

Regression Coefficients across Different Streams and Data Granularity Scenarios

In this study, we used Andrews' (2019) linear regression models to predict temperature gradients in Michigan's streams. There were two main advantages of using these statistical models. First, these statistical models did not require complex mathematical calculations and extensive datasets. This is an important feature of statistical models because they make complex environmental variables (e.g., shortwave solar radiation) simpler to parameterize to be included in models (Cheng & Wiley, 2016). Moreover, less need for extensive datasets reduces the time, effort and financial resources that must be invested in data collection procedures. Second, regression coefficients clearly revealed the magnitude and type of the relationship between environmental variables and the response variable. To illustrate, if a coefficient had negative sign, then that parameter was conversely related with the

response variable. This information was useful for understanding thermal dynamics in streams and most effective factors that influence the temperature gradient. Furthermore, regression coefficients could be used for testing scenarios that reflect various environmental conditions (e.g., groundwater withdrawal, air temperature) (Caldwell et. al., 2014; Andrews, 2019). Therefore, observing the response of the regression coefficients to different characteristics of streams and data granularity was important to have a better perspective on these linear regression models.

Our results revealed that the coefficient value of the same parameter varied across the streams within the same data granularity. This was an expected outcome considering different characteristics of each stream. For example, a further analysis on stream data revealed that average upstream discharge values varied substantially across streams. For example, average upstream discharge (Q_{up}) ranged between 0.035 m^3/s (Hasler Creek) and 1.618 m^3/s (Carp River) across streams between June-October 2016. Likewise, average upstream temperature (T_{up}) varied between 12.873 °C (Cedar River) and 19.346 °C (Hasler Creek) within the same time period. Therefore, we observed wide range of coefficient values across streams. In other words, each model (e.g., Model 10) was stream-specific even though all model parameters were commonly applied for all streams. Certainly, this also resulted in different model performances for each stream.

Our results also revealed that the model coefficient values in Model 10 changed across data granularity scenarios for Tobacco River. In other words, the weight of some model parameters on model predictions varied between granularity scenarios. This was a result of lower number of data points and lower variation across these data points that was caused by averaging the observations (Dertli, 2021). Consequently, the weight of each parameter changed across the granularity scenarios. In addition, the sign changes in model coefficient of the same parameter indicated parameter instability, which is an indicator of high levels of multicollinearity (Dertli, 2021). This situation has been addressed by many other studies in literature. For example, Mason & Perreault (1991) concluded that low sample size (e.g., $n = 30$) exacerbated the influence of multicollinearity in multiple regression analysis. Furthermore, Kroll & Song (2013) revealed that the effects of multicollinearity in regression models that were developed with ordinary least squares (OLS) increased with smaller sample size.

The Effect of Data Granularity on Model Fitness

As Model 10 was the most complex regression model with eight environmental parameters, Model 10 had the highest model prediction power in all data granularity scenarios (Table 5; Figure 3). Model 8 and Model 9 were other two models, which had the highest mean adjusted correlation coefficient values. One common feature of all these models was that they had day length (S) (i.e., Model 8) or altitude angle (α) (i.e., Model 9) or both (i.e., Model 10) as predictor variables. Another common feature was that they all had separated heat transfer variables (i.e., ΔT_{up}, ΔT_{base}, and ΔT_{base}) rather than cumulative heat transfer variable (i.e., ΔT_{flow}). Separating cumulative heat transfer variable into three different predictor variables increased the explanatory power of models since each these variables reflects different environmental processes separately. Model 1, Model 2, Model 3 and Model 4 were diverged from the rest of the models as they had significantly lower prediction power compared to other models (Figure 3). None of these models included neither day length (S) nor altitude angle (α). In other words, including at least one of these predictor variables substantially increased model fit. Therefore, we concluded that these variables were very important in temperature gradient predictions. This conclusion was reasonable because these variables were included in the models to reflect the influence of exposure time of streams to the solar radiation, and to illustrate the importance of solar radiation in temperature dynamics in riverine systems, which was addressed in various studies in literature (Dingman, 1972; Sinokrot & Stefan, 1993; Sridhar et. al., 2004; Dugdale et. al., 2018).

In our study, it was clearly shown that higher data granularity resulted in higher overall model fit (Table 5; Figure 3). As stated in Dertli (2021), this might be a consequence of reduced sample size (i.e., number of observations) with the aggregation of observations by taking their average. However, further observations in the same study showed that higher data granularity reduced the model fit for some streams. In other words, higher data granularity does not always result in high model fit. More legitimate reason was unique characteristics of streams that resulted in different outcomes under each data granularity scenarios. For example, the value of $R^2_{adj.}$ of Model 10 for Tobacco River under 12-hours granularity scenario was lower when compared to the same value under 6-hours data granularity scenario (Dertli, 2021). However, in the same study, it was shown that the value of $R^2_{adj.}$ of Model 10 for Carp River under 12-hours data granularity was higher than the same value under daily data granularity. To draw a better picture of the variations between stream characteristics, we obtained average downstream stream temperature and average downstream discharge values of each stream as provided in Table A1.

As mentioned before, the unique characteristics of streams were already reflected in the responses of model coefficients to changing data granularities. In addition to our findings, Dertli (2021) showed that parameter coefficients for ΔT_{up}, ΔT_{base}, and ΔT_{base} had different responses to data granularity change from hourly to weekly for each stream. This implied that different responses of model parameters caused the variations between increase or decrease patterns of model fit (i.e., $R^2_{adj.}$) across data granularity scenarios for each stream.

All these results showed that the unique characteristics of streams are determining factor of the model fit and they influence response of the model coefficients to different data granularity scenarios. Although overall model fit increased with higher data granularity in overall, it is not possible to propose a universal rule, such as "high data granularity should be used to achieve high model performances". Moreover, selected data granularity may not be useful for answering particular research or environmental management questions, even though the models yield robust predictions. For example, using hourly stream temperature estimates to predict seasonal fish distributions would not be appropriate (due to the temporal scale mismatch) even if model fit is higher with hourly data granularity. Therefore, it is not possible to suggest the "best" data granularity for all modeling approaches in environmental management practices. However, arbitrary selection of data granularity should be avoided because it can have consequences in model-based decision-making processes in environmental sciences. Since the ecological relevance of data granularity should be as important as the model prediction power, regression models should be designed to have the highest model fit with the most ecologically-relevant data granularity.

The Effect of Data Granularity on Model Selection

Model weight was a useful indicator of the level of model complexity-model fit balance. High model weight (i.e., maximum value of 1) of a model was an indicator of high model fit with the minimal number of explanatory variables compared to the other models that were included in the model weight analysis. The percentages that are shown in the results (Figure 4) indicated the proportion of the total number of streams (n = 16) for which a model had the highest model weight. For example, Model 10 had the highest model weight for 62.5% of all streams, while Model 11 had the highest model weight for 25% of all streams. This revealed the best possible model selection for each data granularity scenario by taking all streams into account. No model had the highest model weight for all streams in all data granularity scenarios. For example, Model 8, which did not include altitude angle (α), had the highest model weight for 6.25% of streams (n = 1) with hourly data granularity. On the other hand, Model 9, lacking day length (S) parameter, had the

highest model weight for another stream within the same data granularity scenario (Table 6). In other words, a particular environmental variable (e.g., day length) may be an important determinant of model predictions for some streams, while it may not be for the other streams. This, again, highlighted the importance of the stream characteristics on model evaluation.

Despite the fact that Model 10 had the highest model weight for the majority of streams in general, increasing data granularity resulted to some changes in model selection results, such that, less complex models (e.g., Model 1, Model 5, Model 6, Model 7) appeared to have highest model weight for more streams with higher data granularity scenarios (i.e., daily and weekly). In other words, the influence of model complexity on model fit may have decreased with higher data granularity. This conclusion was congruent with the relationship between model predictive power and data granularity. Since model predictive power generally increased with higher data granularity, higher model predictive powers were achieved with a smaller number of model parameters. This conclusion implied that less complex models may be more useful and efficient to predict response variables for higher data granularities. For example, Arismendi et. al. (2014) evaluated stream temperature predictions simple linear regression model that only included regional air temperature. They averaged daily air temperatures into weekly air temperatures, and they found that their model had an average Nash–Sutcliffe efficiency (NSE) value of 0.86. Although NSE and adjusted correlation coefficient ($R^2_{adj.}$) use slightly different methods to evaluate model fit, explaining 86% of the variation between observed and predicted values can be considered a significantly high model performance for such simple model. Therefore, selection of high data granularities in data may be advantageous since it may allow modelers to adopt simple models for environmental predictions. By using such simple models, researchers may avoid dealing with complicated models, extensive data collection requirements, and possible effects of multicollinearity.

Conclusions

1. Selection of data granularity can affect the model coefficients (both magnitude and sign). This may result in biases in interpretation of environmental variables, and consequently can lead to a mismanagement of ecosystems. In addition, the chance of having multicollinearity in models can increase with higher data granularity. Multicollinearity can also cause misinterpretations of environmental variables especially when parameter instability in the model coefficients occur.

2. Model fitness may be affected by data granularity selection, which may lead to misevaluation of models. Moreover, characteristics of the streams

determine the influence of higher data granularity on model prediction power. For some streams, higher data granularity increases the model fitness while it reduces model fitness for other streams. Therefore, it is not possible to conclude that higher data granularity certainly results in higher model fitness. Although we did not address the issue in our study, selection of time period (e.g., July data) may also potentially influence the relationship between data granularity and model prediction power (Dertli, 2021).

3. Selection of best models based on the rule of parsimony may be influenced by the selection of data granularity. Since higher data granularity decreases the number of data points, it can make simpler models better predictors. In addition, selection of data granularity may change the significance of environmental variables in model parsimony. Therefore, data granularity selection is important for model designing processes.

4. Certainly, our study did not propose such data granularity type that should be used to obtain high model robustness in general sense. However, we have shown that evaluation of model fit, model selection and interpretation of the model results and environmental variables can substantially vary with data granularity selection. Therefore, we highly recommend researchers avoid arbitrary choice of data granularity and make data granularity selections based upon their relevance to their research and management purposes.

Acknowledgement

The authors thank Ryan Andrews for designing the statistical models that were used in this research and for the efforts in data collection processes. The authors also thank Dr. Scott Peacor for contribution on this research. In addition, the authors thank Michigan Department of Natural Resources, Michigan State University Department of Fisheries and Wildlife, and Republic of Turkey Ministry of Agriculture and Forestry for funding this study.

References

Akaike, H. (1973). Maximum likelihood identification of Gaussian autoregressive moving average models. *Biometrika, 60*(2)*, 255-265*. **https://doi.org/10.1093/biomet/60.2.255**

Akossou, A. Y. J., & Palm, R. (2013). Impact of data structure on the estimators R-square and adjusted R-square in linear regression. *International Journal of Mathematics and Computation. 20*(3)*,* 84-93. **https://www.researchgate.net/publication/289526309_Impact_of_data_structure_on_the_estimators_R-square_and_adjusted_R-square_in_linear_regression**

Andrews, R. (2019). *Effects of flow reduction on thermal dynamics of streams: improving an important link in Michigan's water withdrawal assessment tool* (Publication No…) [Master's thesis, Michigan State University]. East Lansing, MI.

Arismendi, I., Safeeq, M., Dunham, J. B., & Johnson, S. L. (2014). Can air temperature be used to project influences of climate change on stream temperature? *Environmental Research Letters, 9*(8), 1-12. **https://doi.org/10.1088/1748-9326/9/8/084015**

Bender, R. (2009). Introduction to the use of regression models in epidemiology. In: Verma M. (Eds.), *Cancer Epidemiology. Methods in Molecular Biology* (vol 471, pp. 179-195). Humana Press.

Caldwell, P., Segura, C., Gull Laird, S., Sun, G., Mcnulty, S. G., Sandercock, M., Boggs, J., & Vose, J. M. (2014). Short-term stream water temperature observations permit rapid assessment of potential climate change impacts. *Hydrological Processes, 29*(9), 2196-2211. **https://doi.org/10.1002/hyp.10358**

Chang, H., Watson, E., & Strecker, A. (2018). Climate change and stream temperature in the willamette river basin: implications for fish habitat. In H.S. Jung & B. Wang (Eds.), *World Scientific Series on Asia-Pacific Weather and Climate: Bridging Science and Policy Implication for Managing Climate Extremes* (pp. 119-132). APCC and World Scientific. **https://doi.org/10.1142/9789813235663_0008**

Chen, Y. D., McCutcheon, S. C., Norton, D. J., & Nutter, W. L. (1998). Stream temperature simulation of forested riparian areas: II. Model application. *Journal of Environmental Engineering,124*(4). **https://doi.org/10.1061/(asce)0733-9372(1998)124:4(316)**

Cheng, S. T., & Wiley, M. J. (2016). A Reduced Parameter Stream Temperature Model (RPSTM) for basin-wide simulations. *Environmental Modeling and Software, 82,* 295-307. **https://doi.org/10.1016/j.envsoft.2016.04.015**

Dertli, H. I. (2021). *The Impact of Data Granularity and Stream Classification on Temperature Gradient Modeling in Michigan's Streams* [Master's thesis, Michigan State University]. **https://doi.org/doi:10.25335/g0t8-1q40**

Dingman, S. L. (1972). Equilibrium temperatures of water surfaces as related to air temperature and solar radiation. *Water Resources Research, 8*(1), 42-49. **https://doi.org/10.1029/WR008i001p00042**

Du, X., Goss, G., & Faramarzi, M. (2020). Impacts of hydrological processes on stream temperature in a cold region watershed based on the SWAT equilibrium temperature model. *Water (Switzerland), 12*(4)*,* 1112. **https://doi.org/10.3390/W12041112**

Ducharne, A. (2008). Importance of stream temperature to climate change impact on water quality. *Hydrology and Earth System Sciences, 12*(3)*,* 797-810. **https://doi.org/10.5194/hess-12-797-2008**

Dugdale, S. J., Malcolm, I. A., Kantola, K., & Hannah, D. M. (2018). Stream temperature under contrasting riparian forest cover: Understanding thermal dynamics and heat exchange processes. *Science of the Total Environment, 610-611,* 1375-1389. **https://doi.org/10.1016/j.scitotenv.2017.08.198**

Ficklin, D. L., Stewart, I. T., & Maurer, E. P. (2013). Effects of climate change on stream temperature, dissolved oxygen, and sediment concentration in the Sierra Nevada in California. *Water Resources Research, 49*(5), 2765-2782. **https://doi.org/10.1002/wrcr.20248**

Guo, D., Lintern, A., Webb, J. A., Ryu, D., Liu, S., Bende-Michl, U.,Leahy, P., Wilson, P., & Western, A. W. (2019). Key factors affecting temporal variability in stream water quality. *Water Resources Research, 55*(1), 112-129. **https://doi.org/10.1029/2018WR023370**

Hamid, A., Bhat, S. U., & Jehangir, A. (2020). Local determinants influencing stream water quality. *Applied Water Science, 10*(1), 1-16. **https://doi.org/10.1007/s13201-019-1043-4**

Iversen, T.M. (1971). The ecology of a mosquito population (A~des communis) in a temporary pool in a Danish beech wood. *Archiv fur Hydrobiologie, 69*, 309-332.

Jackson, H. M., Gibbins, C. N., & Soulsby, C. (2007). Role of discharge and temperature variation in determining invertebrate community structure in a regulated river. *River Research and Applications, 23*(6), 651-669. **https://doi.org/10.1002/rra.1006**

Kroll, C. N., & Song, P. (2013). Impact of multicollinearity on small sample hydrologic regression models. *Water Resources Research, 49*(6)*,* 3756-3769. **https://doi.org/10.1002/wrcr.20315**

Magnusson, J., Jonas, T., & Kirchner, J. W. (2012). Temperature dynamics of a proglacial stream: Identifying dominant energy balance components and inferring spatially integrated hydraulic geometry. *Water Resources Research, 48*(6), 1-16. **https://doi.org/10.1029/2011WR011378**

Mantua, N., Tohver, I., & Hamlet, A. (2010). Climate change impacts on streamflow extremes and summertime stream temperature and their possible consequences for freshwater salmon habitat in Washington State. *Climatic Change, 102,* 187-223. **https://doi.org/10.1007/s10584-010-9845-2**

Mason, C. H., & Perreault, W. D. (1991). Collinearity, Power, and Interpretation of Multiple Regression Analysis. *Journal of Marketing Research, 28*(3), 268-280. **https://doi.org/10.2307/3172863**

Neumann, D. W., Rajagopalan, B., & Zagona, E. A. (2003). Regression Model for Daily Maximum Stream Temperature. *Journal of Environmental Engineering, 129*(7). **https://doi.org/10.1061/(asce)0733-9372(2003)129:7(667)**

Nuhfer, A. J., Zorn, T. G., & Wills, T. C. (2017). Effects of reduced summer flows on the brook trout population and temperatures of a groundwater-influenced stream. *Ecology of Freshwater Fish, 26*(1), 108-119. **https://doi.org/10.1111/eff.12259**

Pilgrim, J. M., Fang, X., & Stefan, H. G. (1998). Stream temperature correlations with air temperatures in Minnesota: Implications for climate warming. *Journal of the American Water Resources Association, 34*(5), 1109-1121. **https://onlinelibrary.wiley.com/doi/10.1111/j.1752-1688.1998.tb04158.x**

Seber, G. A. F., & Wild, C. J. (1989). Autocorrelated Errors. In *Nonlinear Regression*. **https://doi.org/10.1002/0471725315.ch6**

Sinokrot, B. A., & Stefan, H. G. (1993). Stream temperature dynamics: Measurements and modeling. *Water Resources Research, 29*(7), 2299-2312. **https://doi.org/10.1029/93WR00540**

Sridhar, V., Sansone, A. L., LaMarche, J., Dubin, T., & Lettenmaier, D. P. (2004). Prediction of stream temperature in forested watersheds. *Journal of the American Water Resources Association, 40(1),* 197-213. **https://doi.org/10.1111/j.1752-1688.2004.tb01019.x**

Stefan, H. G., & Preud'homme, E. B. (1993). Stream temperature estimation from air temperature. *Journal of the American Water Resources Association, 29*(1), 27-45**.** **https://doi.org/10.1111/j.1752-1688.1993.tb01502.x**

Webb, B. W., Clack, P. D., & Walling, D. E. (2003). Water-air temperature relationships in a Devon river system and the role of flow. *Hydrological Processes, 17*(5), 3069-3084. **https://doi.org/10.1002/hyp.1280**

Zorn, T.G., Seelbach, P.W., & Wiley, M.J. (2004). *Utility of Species-Specific, Multiple Linear Regression Models for Prediction of Fish Assemblages in Rivers of Michigan's Lower Peninsula*. Michigan Department of Natural Resources. **http://www.michigandnr.com/PUBLICATIONS/PDFS/ifr/ifrlibra/Research/reports/2072rr.pdf**

Zorn, T. G., Seelbach. P. W., Rutherford, E. S., Wills,T. C., Cheng, S., & Wiley, M. J. (2008). *A landscape-scale habitat suitability model to evaluate effects of flow reduction on fish assemblages in Michigan streams*. Michigan Department of Natural Resources. **https://www2.dnr.state.mi.us/Publications/pdfs/ifr/ifrlibra/Research/reports/2089/RR2089.pdf**

Appendix

Table A1

Observed Average Stream Temperature and Average Discharge Values of Streams between June-October in 2016

Stream	$\overline{T_w}$	$\overline{T_{down}}$	$\overline{Q_{up}}$	$\overline{Q_{down}}$
BL	15.148	15.412	0.830	0.727
CR	17.286	17.269	1.618	1.510
CC	14.371	14.408	0.562	0.443
MG	18.232	17.780	0.136	0.150
PK	17.107	17.481	0.477	0.551
BF	16.147	15.300	0.127	0.221
CP	17.286	17.269	1.618	1.510
PG	17.223	16.837	0.503	0.667
SP	17.550	17.452	0.190	0.287
ESC	17.476	17.424	0.986	1.268
NTW	21.190	20.355	0.790	0.259
TB	16.350	16.905	0.518	0.526
HS	19.286	17.835	0.035	0.089
PR	17.477	17.813	0.287	0.323
SQ	16.247	17.207	0.036	0.122
SW	19.234	19.511	0.453	0.093

Note. $\overline{T_w}$: average upstream temperature (℃), $\overline{T_{down}}$: average downstream temperature (℃), $\overline{Q_{up}}$: average upstream discharge (m^3/s), $\overline{Q_{down}}$: average downstream discharge (m^3/s).

Extended Turkish Abstract
(Genişletilmiş Türkçe Özet)

Veri Taneselliğinin Michigan Akarsularının Sıcaklık Gradyan Modellemesi Üzerindeki Etkisi

Akarsu sıcaklıklarının tatlı su ekosistemlerindeki fiziksel, kimyasal ve biyolojik süreçlerde önemli bir rolü bulunmaktadır. Bu sebeple akarsu sıcaklığının, su kalitesi ve ekosistem işlevselliğinin önemli parametrelerinden birisi olduğu düşünülmektedir (Guo ve ark., 2019; Hamid ve ark., 2020). Su sıcaklığı, birçok farklı türün (ör. birincil üreticiler, sucuk omurgalı ve omurgasızlar) hayatta kalma, üreme ve yayılma süreçlerini etkilediği için, su sıcaklıklarını belirleyen faktörlerin anlaşılması kritik bir önem arz etmektedir (Iversen, 1971; Jackson ve ark., 2007; Zorn ve ark., 2004; Nuhfer ve ark., 2017). Literatürde akarsu sıcaklıklarını belirleyen farklı meteorolojik (örn. hava sıcaklığı) ve hidrolojik (örn. yağış) süreçler ele alınmıştır (Du ve ark., 2020). Ancak bu ekosistemlerin açık sistemler olması, dolayısıyla bu süreçlerin birbiriyle de etkileşmesi akarsu sıcaklıklarına etki eden faktörlerin anlaşılmasını zorlaştırmaktadır. Bu noktada, istatistiksel modeller araştırmacılara karmaşık sistemlerin açıklanmasında yardımcı olmaktadır.

İstatiksel modeller literatürde doğal süreçlerin açıklanmasında sıkça kullanılmaktadır. Örneğin, regresyon modellerinin değişken katsayıları sayesinde bu süreçlerin birbirleriyle etkileşimi matematiksel olarak açıklanabilmektedir (Bender, 2009). Ayrıca bu sayede çevresel değişimlerin (örn. iklim değişikliği, yeraltı sularının çekilmesi) akarsu termodinamiğine etkileri tahmin edilebilmektedir (Mantua, & Tohver, 2010; Andrews, 2019). Ancak istatistiksel modeller çevresel araştırmalar için çok önemli olsa da bu modellerden alınacak çıktılar kullanılan verinin yapısına oldukça bağlı olabilmektedir (Akossou, & Palm, 2013). Örneğin, veri taneselliği verideki gözlem sayısını, çoklu doğrusal bağlantı (multicollinearity) miktarını ve model çıktılarını etkileyebildiğinden verinin önemli yapısal özelliklerinden sayılmaktadır (Dertli, 2021). Ancak birçok çalışmada, modellerde kullanılan veri taneselliği verinin ekolojik anlamına uygun olarak seçilmesine rağmen, diğer birçok çalışmada kullanılan veri taneselliği keyfi olarak seçilmektedir ya da bu seçimin sebebi açıklanmamaktadır. Bu keyfi seçim, kullanılan modellerin başarısının değerlendirilmesinde ve model çıktılarının yorumlanmasında yanılgılara sebep olabilmektedir (Stefan, & Preud'homme, 1993; Pilgrim ve ark., 1998; Webb ve ark. 2003).

Bu çalışmada, kullanılan veri taneselliğinin doğrusal regresyon modelleri ve bu modellerin yorumlanmasındaki etkisi ele alınmıştır. Bu çalışmada amaçlanan hedefler:

1. Regresyon değişken katsayılarının farklı veri taneselliği senaryolarında değerlendirilmesi ve yorumlanması,
2. Regresyon model uyumunun (fitness) değişiminin farklı veri taneselliği senaryolarında değerlendirilmesi ve yorumlanması,
3. Parsimoni ilkesine bağlı olarak model seçiminin farklı veri taneselliği senaryolarında değerlendirilmesi ve yorumlanması,

olarak belirlenmiştir. Bu hedefler doğrultusunda, veri taneselliğinin regresyon modelleri ve çıktıları üzerine ki etkilerinin ayrıntılı bir şekilde analiz edilmesi amaçlanmıştır.

Bu çalışmada, Michigan'da farklı bölgelerde bulunan 16 akarsudan elde edilen veriler, Andrews (2019) tarafından geliştirilen regresyon modelleri üzerinde, su sıcaklığı değişiminin (sıcaklık gradyan, ΔT) (℃) tahmin edilmesi amacıyla seçilmiş ve kullanılmıştır. Andrews (2019) tarafından geliştirilen bu 11 model, hiyerarşik model geliştirme yöntemi ile her adımda modele yeni parametreler eklenerek dizayn edilmiştir. Bu akarsulara ait verilerde bulunan gözlemlerin ortalaması farklı zaman dilimlerine göre alınıp, 1-saat, 2-saat, 6-saat, 12-saat, 24 saat, 168-saat olmak üzere, 6 farklı veri taneselliği senaryosu elde

edilmiştir. Bu farklı senaryolara sahip veriler regresyon modellerinde yürütülerek model çıktıları elde edilmiştir. Regresyon katsayıları bu simülasyonlar neticesinde elde edilmiştir. Model uyumu modellerin çıktılarının gözlemlerle olan korelasyon miktarına bağlı olarak değerlendirilmiştir. Farklı veri taneselliği senaryolarının model seçimde neden olduğu değişiklikler Akaike Bilgi Kriteri (Akaike's Information Criterion-AIC) değerleri kullanılarak elde edilen model ağırlıklarına bağlı olarak değerlendirilmiştir (Akaike, 1973).

Regresyon katsayı analizi iki önemli bulguyu ortaya çıkarmıştır. Birincisi, Model 10 1-saat veri taneselliği senaryosunda bütün akarsular için yürütüldüğünde aynı model parametre katsayılarının değerlerinin (örn. hava-su sıcaklık farkı ($T_a - T_w$)) akarsular arasında nicelik ve nitelik olarak değişmiştir. İkincisi, aynı akarsuya ait veride farklı veri taneselliği senaryoları kullanıldığında, Model 10'a ait parameter katsayılarının nicelik ve nitelik olarak değiştigi gözlenmiştir. Örneğin, hava-su sıcaklık farkı ($T_a - T_w$) parametre katsayısının değeri 1–saat senaryosunda 0.023 olarak ölçülürken, 24-saat senaryosunda -0.023 olarak ölçülmüştür. Model uyumluluk analizleri de önemli bulgular ortaya koymuştur. Örneğin, Model 10'a hava korelasyon değerlerinin (ya da uyumluluğunun) diğer modellerden daha yüksek olduğu gözlenmiştir. Model uyumluluğu analizi ayrıca veri taneselliğinin artışının (1-saat'lik senaryodan 168-saat'lik senaryoya) genel olarak model uyumluluğunu (fitness) artırdığını göstermiştir. Parsimoni ilkesine bağlı olarak model seçimi analizleri, Model 10'un diğer modellere kıyasla daha fazla sayıda akarsu (akarsuların %62,5'i) için daha iyi çalıştığı gözlemlenmiştir. Bununla beraber, veri taneselliği bu seçimlerde değişikliğe neden olmuştur. Örneğin, 168-saat senaryosunda Model 10 sadece akarsuların %31.25'i için diğer modellere kıyasla daha iyi çalıştığı gözlemlenmiştir.

Sonuç olarak bu çalışma regresyon modelleri ve kullanılan veri yapısı açısından bazı önemli sonuçlar ortaya koymuştur. Örneğin, regresyon modellerdeki parametre katsayılarının nicelik ve niteliğinin kullanılan veri taneselliğine bağlı olarak değişebileceğinin gösterilmesi, bu modellerde bulunan parametrelerin model çıktıları üzerindeki etkisinin de taneselliğe bağlı olarak değişebileceği gösterilmiştir. Bu sonuç, modellerde kullanılan parametrelerin (ör. $T_a - T_w$) veri taneselliğine bağlı olarak akarsu sıcaklığı tahminlerine olumlu ya da olumsuz olarak etki edebileceğini göstermiştir. Bu durum, modellerde uygun olmayan bir veri taneselliği kullanıldığında akarsu sıcaklığına etki eden faktörlerin yanlış yorumlanmasına sebep olabileceğini göstermiştir. Ayrıca kullanılan veri taneselliğinin model uyumluluğuna doğrudan etki etmesi, bu modellerin uyumluluklarının değerlendirilmesinde yanlış yorumlamalara sebebiyet verebileceği ortaya koyulmuştur. Bu durum aslında uyumluluğu yüksek olan bir modelin, modele uygun olmayan bir veri taneselliği kullanıldığında uyumluluğunun düşük ölçülebileceğini göstermiştir. Bununla birlikte, kullanılan veri taneselliği, bir akarsu için kullanılması en uygun olan modelin seçimini etkileyebileceğinden, veri taneselliği seçiminin model seçimlerinde yanlış kararlara yol açabileceği sonucu ortaya çıkmıştır.

Şuna dikkat çekmek gerekir ki bu çalışma hangi veri taneselliğinin daha iyi olduğunu ortaya koymayı amaçlamamıştır. Çünkü veri taneselliğinin artması ya da azalması her model veya her durum için farklı sonuçlar ortaya çıkarmaktadır. Bu çalışmanın asıl amacı, keyfi olarak seçilen veri taneselliğinin modeller ve model yorumlamaları üzerindeki muhtemel etkilerine dikkat çekmek ve modellemecilere daha geniş bir bakış açısı sunmaktır. Modellerde kullanılan veri taneselliğinin keyfi olarak değil, araştırmanın cevap bulmaya çalıştığı sorulara uygun olarak seçilmesi, bu modellerin uygunluğunun ve başarısının objektif bir biçimde değerlendirilmesinde büyük önem teşkil edecektir.