Matsuzaki, K.M., Suzuki, N., and Tada, R., 2020, An intensified East Asian winter monsoon in the Japan Sea between 7.9 and 6.6 Ma: Geology, v. 48, https://doi.org/10.1130/G47393.1

SUPPLEMENT FOR AN INTENSIFIED EAST ASIAN WINTER MONSOON IN THE JAPAN SEA BETWEEN 7.9 AND 6.6 MA

K. M. Matsuzaki[1], N. Suzuki[2], R. Tada[3]

1: Atmosphere and Ocean Research Institute, The university of Tokyo, 5-1-5 Kashiwanoha, Kashiwa, Chiba 277-8564, Japan.

2: Department of Earth Science, Graduate School of Science, Tohoku University, Aramaki Aza Aoba 6-3, Aoba-ku, Sendai 980-8578, Japan.

3: Department of Earth and Planetary Science, Graduate School of Science, The University of Tokyo, 7-3-1, Hongo, Bunkyo-ku, Tokyo 113-0033, Japan.

To estimate radiolarian-based SSTs during the late Miocene, the usual models developed for the Pleistocene are not suitable because only a few late Miocene species are extant in the modern ocean. For the purpose of examining long-term paleoceanographic changes of the Japan Sea since 10 Ma, Matsuzaki et al. (2018) selected 14 extant radiolarian species groups whose geographic distribution during the late Miocene was close to that at present (for details see Matsuzaki et al., 2018) (Table S1).

Here we propose to estimate SSTs in the Japan Sea by using the extant species identified by Matsuzaki et al. (2018) conducting a sophisticated multiple regression analysis. For successful and appropriate SST estimation, a modern calibration dataset is needed and a statistical analysis of the data consisting of several steps needs to be conducted. In this study, we used the calibration dataset of Matsuzaki and Itaki (2017), who analyzed changes in the radiolarian assemblage in surface sediment samples collected from low to high latitudes of the northwest Pacific and calibrated their assemblages to SSTs provided by the World Ocean Atlas 2013 (Locarnini et al., 2013) (Table S1) (for details see Matsuzaki and Itaki, 2017, and references therein). The statistical analyses conducted prior to the multiple regression analysis are explained below.

In the first step, we checked whether SST and the selected species (objective and explanatory variables) have a Gaussian distribution in the northwest Pacific by performing a normal quantile-quantile plot (Q-Q plot) (Figs. S1 and S2). On this plot, almost all of the SSTs are within the 95% confidence interval (CI) of the theoretical Gaussian distribution line; thus, SST can be considered to have a Gaussian distribution in the northwest Pacific (Figure S1).

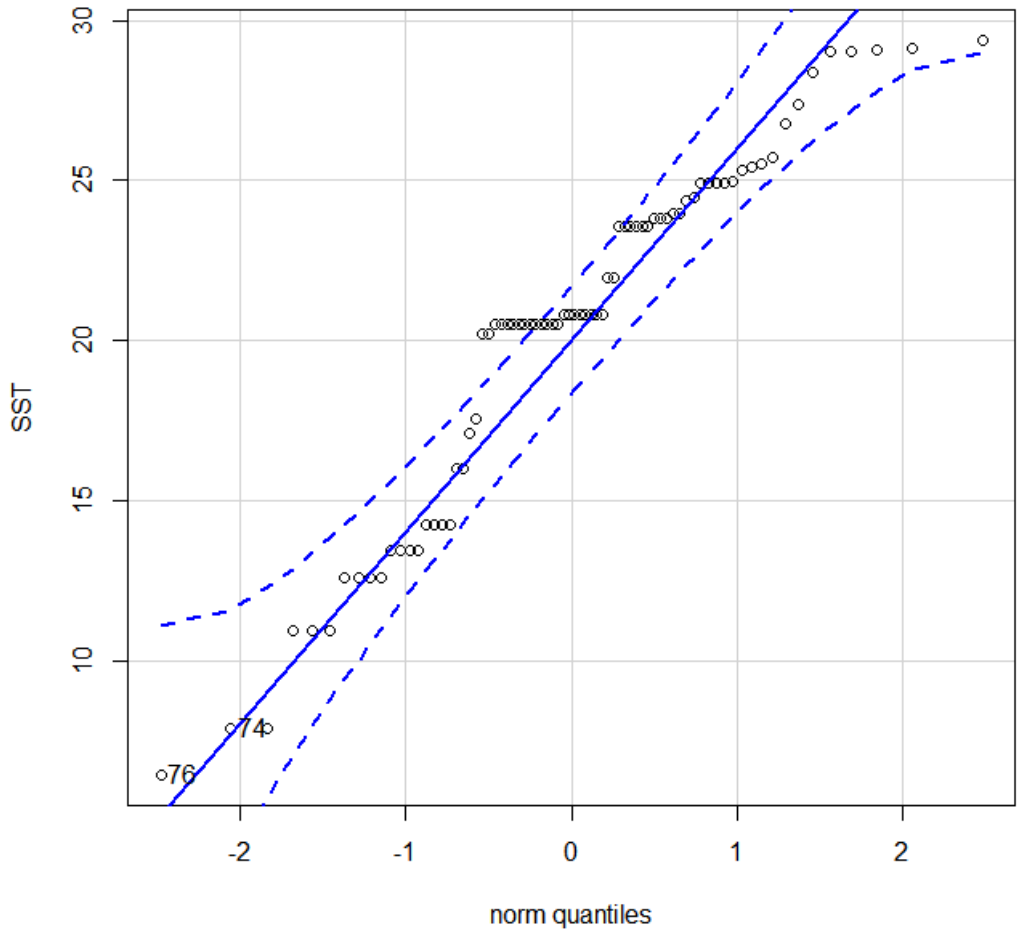
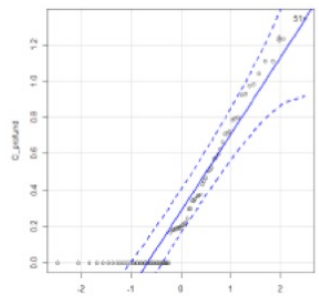
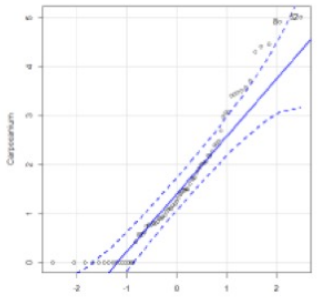


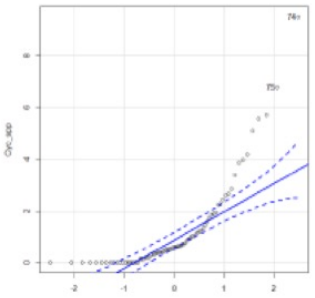
Figure S1. Q-Q plot of annual SSTs, based on samples analyzed and calibrated by Matsuzaki and Itaki (2017) (temperatures from World Ocean Atlas 2013; Locarnini et al., 2013). The solid blue line corresponds to the theoretical Gaussian distribution line. The dotted blue lines indicate the 95% confidence interval (95% CI), automatically created by the R software.



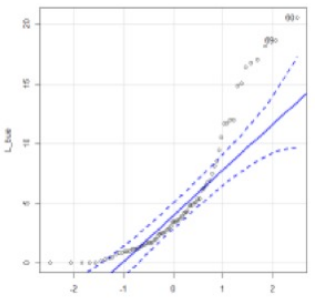
Q-Q plot for *Cornutella profunda* grp.



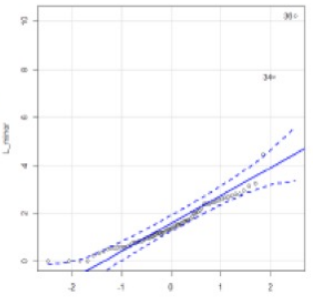
Q-Q plot for *Carpocanistrum* spp.



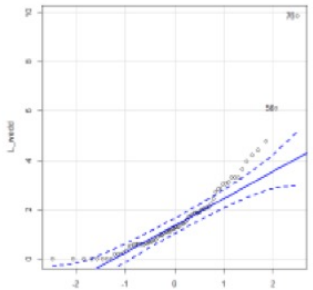
Q-Q plot for *Cycladophora* spp.



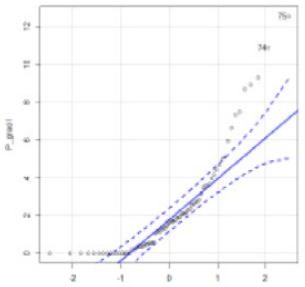
Q-Q plot for *Larcopyle buetschlii* grp.



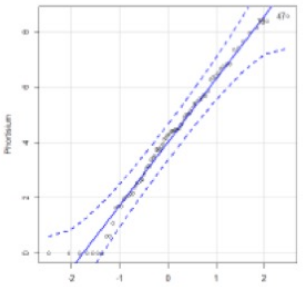
Q-Q plot for *Lithelius minor* grp.



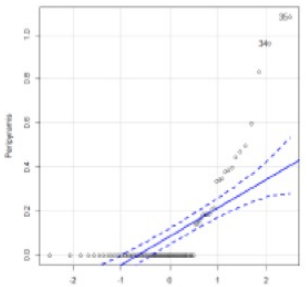
Q-Q plot for *Larcopyle weddellium* grp.



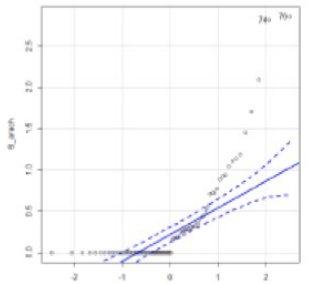
Q-Q plot for *Pseudodictyophimus gracilipes* grp.



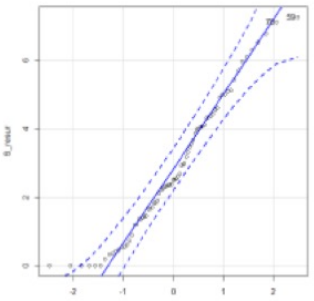
Q-Q plot Peripyramis spp.



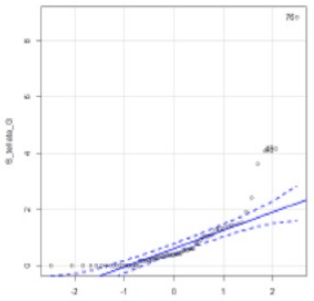
Q-Q plot *Phorticium* spp.



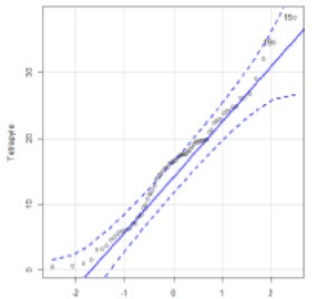
Q-Q plot *Shiphocampe arachnea* grp.



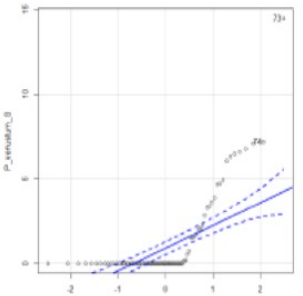
Q-Q plot for *Spongodiscus resurgens* grp.



Q-Q plot for *Stylodictya stellata/ tenuispina* grp.



Q-Q plot *Tetrapyle* spp.



Q-Q plot *Perichlamydium venustum* Grp.

Figure S2. Q-Q plots for the 14 species groups identified as candidate explanatory variables in this study. The solid blue lines correspond to the theoretical Gaussian distribution line, and the dotted blue lines indicate the 95% CI, automatically created by the R software.

Multiple regression analysis is based on the assumption that each explanatory variable follows a Gaussian distribution. Therefore, we also used Q-Q plots to check whether the calibration dataset of 14 species groups selected as candidate explanatory variables have Gaussian distributions (Fig. S2 and Table S2). Several species groups deviated moderately from a Gaussian distribution: *Cycladophora* spp., *Larcopyle buetschlii* group, *Pseudodictyophimus gracilipes* group, *Siphocampe arachnea* group, *Carpocanistrum* spp., and *Larcopyle weddellium*. These data thus needed to be transformed. In addition, the *Stylochlamydium venustum* group deviated strongly from a normal distribution and was therefore excluded from further analysis (Fig. S2 and Table S2).

We conducted Box–Cox transformations of the relative abundances of *Cycladophora* spp., *L. buetschlii* group, *P. gracilipes* group, *S. arachnea* group, *Carpocanium* spp., and *L. weddellium* based on Hair et al. (2014) as follows:

$$*= \begin{cases} \frac{(y+1)^{\lambda}-1}{\lambda}, & \lambda \neq 0 \\ \log(y+1), & \lambda = 0 \end{cases},$$

where λ = 1/2 corresponds to a square root transformation; λ = 0 corresponds to a log transformation; and λ = –1/2 corresponds to an inverse square root transformation.

With λ = 0, *Cycladophora* spp., *L. buetschlii*, *S. arachnea* group, *Carpocanium* spp., and *L. weddellium* were more normally distributed, whereas with λ = –1/2, the *P. gracilipes* group and *Stylodictya stellata* group were more normally distributed.

The selected explanatory variables should be related to SST for suitable SST estimation by multiple regression analysis. In the next step, we therefore checked whether the 13 retained species having a Gaussian distribution were related to SST. For species whose changes in relative abundances were not related to changes in SSTs, we performed a log transformation of the data. If the log-transformed relative abundances of those species did not show a relation with changes in SST, we conducted a Box–Cox transformation of the log-transformed relative abundances with λ = –1/2. If after the Box–Cox transformation there was still no clear relation with SST, we excluded the species group from the multiple regression analysis. Following these procedures, *Spongodiscus resurgens*, *Phorticium* spp., *Tetrapyle* spp., *L. minor* group, and *Carpocanium* spp. showed a clear correlation with temperature without any transformation; the *L. buestchlii* group showed a clear correlation with SST after a log transformation; and the *S. stellata* and *P. gracilipes* groups showed a correlation with

SST after the Box–Cox transformation with λ = –1/2. The remaining five species groups were discarded because of their poor relationship with SST (Table S3).

Multicollinearity occurs when explanatory variables are related by a linear function; the existence of multicollinearity makes it impossible to estimate the regression coefficient (Everitt and Skrondal, 2002). We can detect multicollinearity of variables by reviewing the variance inflation factor (VIF) and relative correlation coefficient (Lomax and Hahs-Vaughn, 2012). If the correlation coefficient is more than 0.65, and the VIF is more than 5, there is multicollinearity (Lomax and Hahs-Vaughn, 2013). For the candidate explanatory variables, therefore, we constructed elliptical correlation matrix plots and a heat map of the correlation matrix (Fig. S3). The dataset is available in Supplement Table S3.

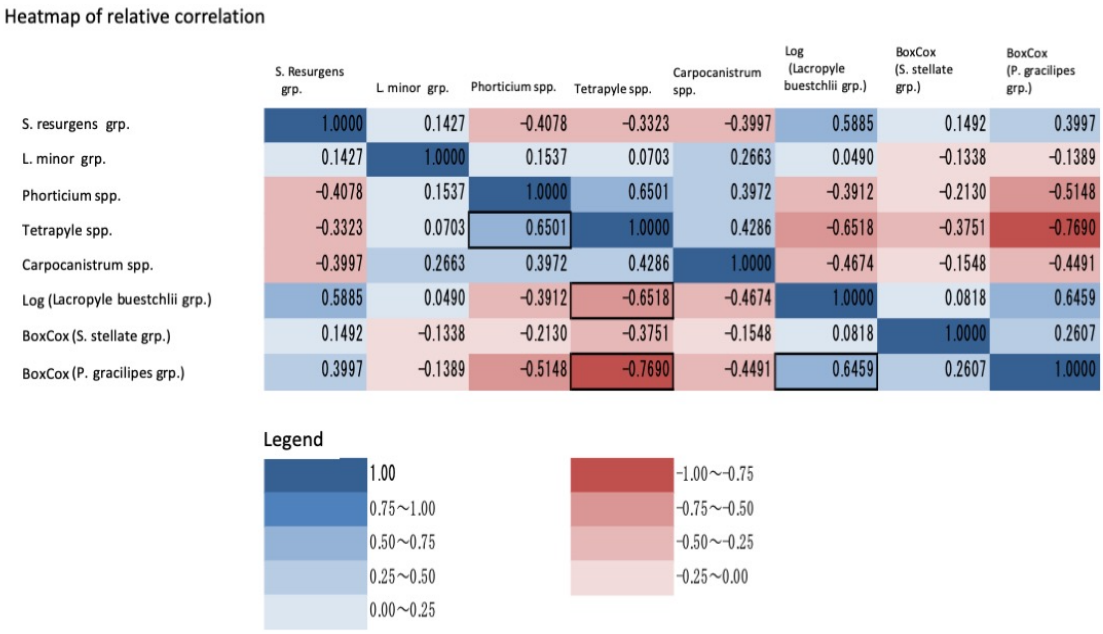


Figure S3. Heatmap of the correlation matrix of eight explanatory variables created by R software

The heatmap shows positive correlations between *Phorticium* spp. and *Tetrapyle* spp. (0.650), and between the *L. buetschlii* group and the Box–Cox converted value of the *P. gracilipes* group (0.646) (λ = –0.5). Therefore, we created two composite variables, one that combined *Phorticium* spp. and *Tetrapyle* spp., and another that combined *L. buetschlii* group with the Box–Cox converted value of the *P. gracilipes* group. The dataset is in Supplement Table S4. We applied the same procedure to the resulting six

variables (the composite variables and the other four variables in original group) to check for any additional multicollinearity, but none was detected.

Residuals:

Min	1Q	Median	3Q	Max
-6.993	-1.3804	0.0294	1.4784	5.398

Residual standard error	2.46	Degree of freedom	67
Multiple R-squared	0.8251	Adjusted R-squared	0.8094
F-statistic:	52.67 on 6 and 67 DF	p-value	2.20E-16

Coefficients:

	Estimated	Standard Error	t-value	Pr(>|t|)	
(Intercept)	21.590	1.938	11.138	0.000	***
S_resur	-0.435	0.183	-2.375	0.020	*
L_minor	0.075	0.345	0.218	0.828	
Comb_Pho_and_Tet	0.177	0.049	3.592	0.001	***
Carpocanium	0.795	0.275	2.895	0.005	**
Comb_Log_bue_and_Box_P_gra	-2.853	0.702	-4.064	0.000	***
Box_Cox..0_5._S_tellata_G	-2.903	0.992	-2.927	0.005	**

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

	Estimated regression coeffiecient	Lower limit of 95% CI	Upper limit of 95% CI	Sandard Error	t-value	p-value
(Intercept)	21.590	17.721	25.459	1.938	11.138	6.79E-17
S_resur	-0.435	-0.800	-0.069	0.183	-2.375	2.04E-02
L_minor	0.075	-0.614	0.764	0.345	0.218	8.28E-01
Comb_Pho_and_Tet	0.177	0.079	0.276	0.049	3.592	6.20E-04
Carpocanium	0.795	0.247	1.343	0.275	2.895	5.11E-03
Comb_Log_bue_and_Box_P_gra	-2.853	-4.253	-1.452	0.702	-4.064	1.29E-04
Box_Cox..0_5._S_tellata_G	-2.903	-4.883	-0.923	0.992	-2.927	4.68E-03

Figure S4. Key statistics of the multiple regression analysis.

After these steps, all of the explanatory variables have a Gaussian distribution. Therefore, we performed a multiple regression analysis in which the number of parameters used in this regression were selected following Akaike (1973), for avoid the risk of underfitting (degree of freedom here, which is 67) (Table S4).

In the final model (Fig. S4), 1Q and 3Q are almost equal and the median is near 0; thus, the conditions of the multiple regression analysis are satisfied. The adjusted R^2 is 0.809, and the standard error of the residuals is 2.46ºC. To check if the model worked well with the data, we performed diagnostic linear regression plots (Fig. S5).

The plots indicate that the proposed multiple regression model is suitable for predict SST for the Mio-Pliocene. The plots of residuals against fitted values indicate that the residuals have a nonlinear distribution. On a Q-Q plot, the residuals are relatively well aligned, indicating that they are normally distributed. The scale-location plot, which shows whether the residuals are spread equally along the range of the predictors, indicates that the error indicated by the variance is relatively uniform. Thus, the diagnostic plots of the residuals tend to indicate that the proposed

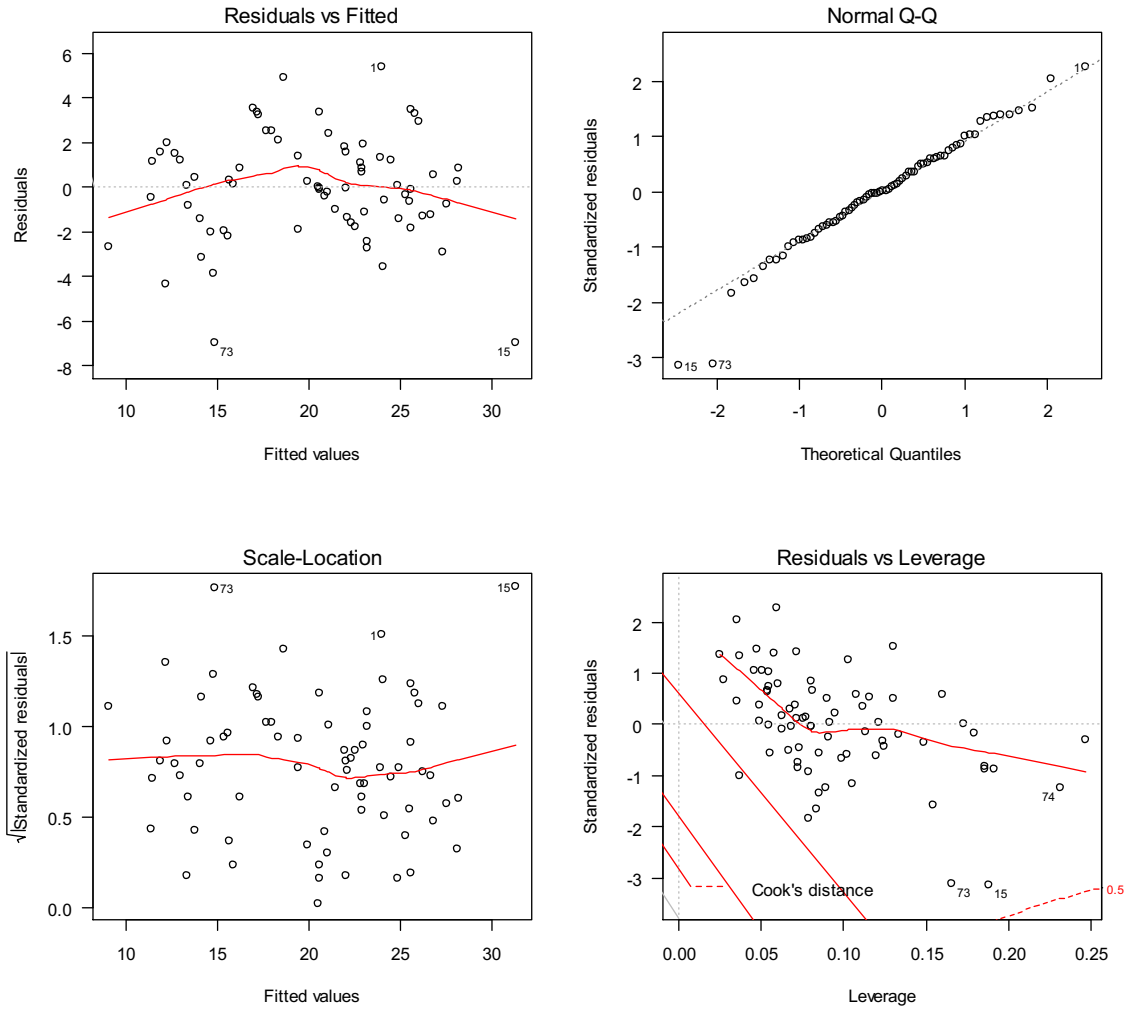


Figure S5. Diagnostic linear regressions to verify the robustness of the selected multiple regression model: residuals vs fitted values; Q-Q plot; scale-location plot, and a residuals vs leverage plot.

multiple regression model is suitable.

Finally, to check whether the model could be improved, we plotted the residuals against leverage. This plot helps us to find outliers that can be removed to improve our model by estimating the Cook's distance of each variable, which indicates how much that variable affects the estimation of the regression coefficient. In our model, samples 15, 73, and 74 from our calibration dataset had a significant effect on the regression coefficient. We can see that these points plot far from the Cook's distance lines; thus, we can potentially obtain an SST estimation model with higher precision SST by removing these three samples. This analysis was carried out with the R software routine "Influence plot".

On the basis of these results, we performed a multiple regression analysis using six variables and excluding samples 15, 73, and 74 (Fig. S6 and Supplement Table S5).

	Estimated regression coeffiecient	Lower limit of 95% CI	Upper limit of 95% CI	Sandard Error	t-value	p-value
(Intercept)	23.310	19.754	26.866	1.780	13.096	9.06E-20
S_resur	-0.275	-0.610	0.061	0.168	-1.636	1.07E-01
L_minor	-0.344	-0.960	0.273	0.309	-1.113	2.70E-01
Comb_Pho_and_Tet	0.142	0.052	0.233	0.045	3.150	2.48E-03
Carpocanium	0.926	0.439	1.413	0.244	3.800	3.25E-04
Comb_Log_bue_and_Box_P_	-3.539	-4.781	-2.297	0.622	-5.691	3.39E-07
Box_Cox..0_5._S_tellata_G	-2.595	-4.376	-0.815	0.891	-2.912	4.94E-03

Residual standard error	2.13	Degree of freedom	64
Multiple R-squared	0.8547	Adjusted R-squared	0.8411
F-statistic:	62.74 on 6 and 64 DF	p-value	2.20E-16

Figure S6. Key statistics of the multiple regression analysis after outlier removal.

The results of the new model are not very different from those of the previous model, but the residual standard error of 2.13°C and the adjusted R^2 of 0.841 are slightly better. Therefore, we the use of diagnostic plots improved our regression model. The final regression equation is as follows:

Annual SST = 23.31 – 0.27 × *S. resurgens* grp – 0.34 × *L. minor* grp + 0.14 × (*Phorticium* spp. + *Tetrapyle* spp.) + 0.93 × *Carpocanium* spp.– 3.54 × [LOG_{10}(1+*Larcopyle buetschlii* grp) + (Box-Cox transformation [λ = –0.5] *P. gracilipes* grp)] – 2.60 × (Box-Cox transformation [λ = –0.5] *S. stellata*/*tenuispina* grp)

According to Matsuzaki et al. (2018), *Spongodiscus resurgens* grp., *Lithelius minor* grp., *Larcopyle buetschlii* grp., and *Pseudodictyophimus gracilipes* grp. are species groups that prefer cold waters; *Stylodictya stellata/tenuispina* grp. prefers temperate-

latitude waters; and *Phorticium* spp. and *Tetrapyle* spp. are associated with warm waters. *Carpocanium* spp. is ambiguous, likely because it inhabits deep waters (Matsuzaki et al., in press). However, according to the dataset of Matsuzaki and Itaki (2017), the highest abundances of this species group is recorded in the tropical North Pacific; thus, we considered this group to be a marker of warm water.

The above regression equation is applicable for temperature from 10°C to 29°C; however, the lower and upper limits of the 95% CI are between 17°C and 25ºC. We performed an F-test (F-statistic) to check the explanatory power of the model and obtained $P = 2.2 \times 10 - 16 < 0.05$; thus, the model has significant explanatory power (Supplement Table S6).

All statistical analyses conducted in this study follow Isomi (2018).

References:

Akaike, H. (1973), Information theory and an extension of the maximum likelihood principle, in: B.N. Petrov and F. Csaki, eds., 2nd lnternat. Syrup. on Information Theory (Akademia Kiado, Budapest) pp. 267 281.

Everitt, B. and Skrondal, A., 2002, The Cambridge dictionary of statistics (Vol. 106). Cambridge: Cambridge University Press.

Hair, J., W. Black, B. Babin, and R. Anderson, 2014, Multivariate Data Analysis: Pearson New International Edition. 7th ed. New Jersey.

Hahs-Vaughn, D.L. and Lomax, R.G., 2013, An introduction to statistical concepts. Routledge.

Isomi, I., 2018, Statistical software "R" for beginners. Data analysis training with test data. Blue Backs. Kodansha, Co. Ltd., Tokyo, 290 p. (in Japanese).

Locarnini, R. A., et al., 2013, World Ocean Atlas 2013, Volume 1: Temperature. S. Levitus, Ed., A. Mishonov Technical Ed.; NOAA Atlas NESDIS 73, 40 pp.

Lomax, R. G., & Hahs-Vaughn, D. L. (2013). An introduction to statistical concepts. Routledge.

Matsuzaki, K. M., and Itaki, T., 2017, New northwest Pacific radiolarian data as a tool to estimate past sea surface and intermediate water temperatures: Paleoceanography, v. 32, p. 218-245, http://doi: 10.1002/2017PA003087.

Matsuzaki, K. M., Itaki, T., Tada, R., and Kamikuri, S. I., 2018, Paleoceanographic history of the Japan Sea over the last 9.5 million years inferred from radiolarian assemblages (IODP expedition 346 sites U1425 and U1430): Progress in Earth and Planetary Science, v. 5, 54, http://doi: 10.1186/s40645-018-0204-7.

Matsuzaki, K. M., Itaki, T., Sugisaki, S. 2020. Polycystine radiolarians vertical distribution in the subtropical Northwest Pacific during Spring 2015 (KS15-4). Paleontological Research, vol. 24, no. 2, pp.1-21.