Supplementary Methods

Description of variance components model

We fitted a model to predict the variance components (VCs) associated with each high frequency measurement (chlorophyll *a* fluorescence, surface water temperature, wind speeds, and PAR) and lake-level attribute (absolute latitude, log absolute longitude, log number of sampling days, log lake area, log maximum lake depth, log mean lake depth, log trophic state, log residence time, water clarity, mean surface water temperature, mean wind speed, and mean PAR). In each lake i, these values could be described by a partially-observed m-dimensional vector $\mathbf{X}_i \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu}$ was a vector of mean values across lakes for each parameter and $\boldsymbol{\Sigma}$ was an estimated covariance matrix. We describe $\mathbf{X}_i$ as partially observed because we measured each of the lake-level attributes but not the VCs. The VCs were instead treated as latent variables that were constrained by observed high-frequency data and sampled from a log-normal multivariate hyper-prior distribution that placed relatively uninformative priors on each hyper-parameter. Hyper-means for each of the VCs in $\boldsymbol{\mu}$ were sampled from a zero-mean normal distribution with standard deviation (SD) of 1, and hyper-standard deviations were sampled from a half-normal zero-mean prior with SD of 5. We then estimated the covariance among the VCs and lake-lake attributes $\boldsymbol{\Sigma}$ by separately modelling the correlation matrix $\boldsymbol{\Omega}$ and diagonal matrix of hyper-standard deviations $\boldsymbol{\Lambda}$ that generate $\boldsymbol{\Sigma} = \boldsymbol{\Lambda\Omega\Lambda}$ (Barnard et al. 2000). The correlation matrix $\boldsymbol{\Omega}$ was given an inverse-Wishart prior of $\sim IW(k+1, \mathbf{I})$. For the lake-level parameters, data were transformed to a common scale with a mean of 0 and SD of 1 for ease estimating $\boldsymbol{\Sigma}$, and we took these values as the true population-level means and SDs in the model, i.e. corresponding elements in $\boldsymbol{\mu}$ and $\boldsymbol{\Lambda}$ were fixed and not estimated.

For Review Only

We constrained the VCs in $\mathbf{X}_i$ by fitting the values to each high-frequency variable y_{ijkl} at each hour j of day k in month l as a function of four coefficients:

$$y_{ijkl} = \alpha_i + \beta_{1il} + \beta_{2ik} + \beta_{3ijkl}, \quad \text{(S1)}$$

where α_i was the estimated mean value of the high-frequency variable in each lake i and β_{1il} and β_{2ik} accounted for the variation among months and days in each lake, respectively. β_{3ijkl} corresponded with variation at the level of hourly observations that was not explained by day and month. We assumed that daily variation was unlikely to vary substantially among months, and so β_{3ijkl} primarily represented an estimate of hourly variance that corresponded with the estimated residual error within each lake. In order to compare the sources of variation in y_{ijkl}, we assumed that values of each of the three temporal coefficients for each of the four high-frequency responses were drawn from independent, zero-mean normal distributions with SDs equal to the estimated VCs:

$$\beta_{i[m]} \sim N(0, \mathbf{X}_{i[m]}),$$

where $\beta_{i[m]}$ was the lake- and time-specific values for the m^{th} β parameter (i.e. corresponding with either β_{ik}, β_{il}, or β_{ijkl}) and $\mathbf{X}_{i[m]}$ was the m^{th} VC for the i^{th} lake. α's were sampled from normal priors with a zero mean and SD of 10.

The observed values were then modelled from a log-normal distribution for chlorophyll *a* fluorescence and normal distribution for water temperature. For wind speeds, Weibull, log-normal, and truncated normal distributions are all widely used for distributions (Carta et al. 2009), but we modelled values from the truncated normal for two main reasons. First, the Weibull and log-normal require wind speeds larger than zero (i.e. positive), which was problematic when measurement resolution was at best 0.001 m s^{-1} and values smaller than this were set to 0. Second, the truncated normal was more intuitive to interpret and has been widely used for predicting wind speeds conditioned on other variables (Hering and Genton 2010; Thorarinsdottir and Gneiting 2010; Baran

For Review Only

2014; Yoder et al. 2014), which were the hour, day, and month of measurement in our case. Finally, we modelled positive PAR values from a mixture of two log-normal distributions, where group membership into one of the two distributions was estimated for each observation. We did this because the data were strongly bimodal, with daytime values much higher than those recorded during night, which approached zero. The probability p_{ijkl} of each observation y_{ijkl} belonging to the first of the two groups was estimated from a Beta distribution that was ~*Beta* (5,5). Membership in the second group could then be derived as $1 - p_{ijkl}$.

Model estimation

We fitted our model within a hierarchical Bayesian framework using Hamiltonian Markov chain Monte Carlo (MCMC) sampling by calling Stan v. 2.16 (Carpenter et al. 2016) from R v. 3.4 (R Development Core Team 2015). Four MCMC chains of 900 iterations were simulated, with a burn-in period of 400 runs. We inferred variance components by calculating posterior means and 95% credible intervals (CI) from a subset of 500 simulations across all four chains. VCs were considered to be different from each other when the 95% CI for the differences between their group means exclude zero.

We used two approaches for assessing model performance in addition to visually inspecting chain traces. First, we calculated the potential scale reduction factor $\hat{R}$ for each parameter from the 500 simulation subset. $\hat{R}$ predicts the extent to which a parameter's credible intervals will be reduced if models are run for an additional number of simulations (i.e., infinite). $\hat{R}$ values were generally less than 1.1, which implied that the model had approximately converged and that MCMC chains had mixed (Gelman and Hill 2007). Second, we ensured that the effective number of simulation draws, n_{eff}, a measure of the independence amongst the subset

For Review Only

of 500 simulations, generally exceeded 100 (Gelman and Hill 2007).

676

References

678 Baran, S. 2014. Probabilistic wind speed forecasting using Bayesian model averaging with truncated normal components. Computational Statistics & Data Analysis **75**: 227–238. 680 doi:10.1016/j.csda.2014.02.013

Barnard, J., R. McCulloch, and X. L. Meng. 2000. Modeling covariance matrices in terms of 682 standard deviations and correlations, with application to shrinkage. Statistica Sinica **10**: 1281–1311.

684 Carpenter, B., A. Gelman, M. Hoffman, and others. 2016. Stan: A probabilistic programming language. Journal of Statistical Software **20**: 1–37.

686 Carta, J. A., P. Ramírez, and S. Velázquez. 2009. A review of wind speed probability distributions used in wind energy analysis: Case studies in the Canary Islands. Renewable and Sustainable 688 Energy Reviews **13**: 933–955. doi:10.1016/j.rser.2008.05.005

Hering, A. S., and M. G. Genton. 2010. Powering up with space-time wind forecasting. Journal of 690 the American Statistical Association **105**: 92–104. doi:10.1198/jasa.2009.ap08117

Gelman, A., and J. Hill. 2007. Data analysis using regression and multilevel/hierarchical models, 692 Cambridge University Press, Cambridge, UK.

R Core Team (2015). R: A language and environment for statistical computing. R Foundation for 694 Statistical Computing, Vienna, Austria.

Thorarinsdottir, T. L., and T. Gneiting. 2010. Probabilistic forecasts of wind speed: ensemble 696 model output statistics by using heteroscedastic censored regression. Journal of the Royal Statistical Society: Series A **173**: 371–388. doi:10.1111/j.1467-985X.2009.00616.x

For Review Only

698 Yoder, M., A. S. Hering, W. C. Navidi, and K. Larson. 2014. Short-term forecasting of categorical changes in wind power with Markov chain models. Wind Energy **17**: 1425–1439.
700 doi:10.1002/we.1641

For Review Only

Table S1. Attributes of 18 lakes sampled for high-frequency chlorophyll fluorescence. Trophic status was calculated as mean annual extracted chlorophyll *a* concentration (collected in the same year that the fluorescence sensor was deployed - asterisks indicate lakes for which only a long-term mean was available), while clarity is given by the diffuse PAR extinction coefficient (K_d – actual or estimated).

Lake	Country	Latitude	Longitude	Data extent (days)	Area (ha)	Max Depth (m)	Mean Depth (m)	Residence time (days)	Trophic status ($\mu g\ L^{-1}$)	Clarity (m^{-1})
Acton	USA	39.55	-84.73	166	244	8	3.5	72	65.3	1.62
Balaton	HUN	46.83	17.73	165	5920	4	3.2	1800	15.7	2.55
Blelham Tarn	UK	54.35	-2.98	363	10	15	6.8	54	21.7	1.6
Douglas	USA	45.57	-84.68	183	1521	27	5.5	1095	2.24	0.32
Feeagh	IRL	53.93	-9.57	341	390	50	14.5	164	1.3	1.79
Harp	CAN	45.37	-79.13	183	71	38	13.3	1168	1.35	0.64
Kentucky	USA	36.73	-88.73	333	59940	17	8	30	15.5	1.48
Lillinonah	USA	41.47	-73.33	188	626	33	13.4	24	30.6*	0.82
Mendota	USA	43.08	-89.37	214	3985	25	12.7	80	5.5	1.37
Müggelsee	DEU	52.43	13.65	218	730	8	4.5	42	44.3	4.06

For Review Only

Ngaroto	NZL	-37.97	175.28	130	117	4	2	65	37.3*	3.12
Paul	USA	46.25	-89.5	98	2	12	3.7	921	2.71	1.21
Taihu	CHN	31.17	120.15	303	22500	3	1.9	264	44.2*	4.4
Tutira	NZL	-39.23	176.92	348	174	42	20.8	730	6.7	0.6
Vanajanselkä	FIN	61.17	24.83	96	10300	24	8	450	13.9	2.24
Vedsted	DNK	55.17	9.33	203	9	11	5	985	61.45	0.83
Võrtsjärv	EST	58.28	26.03	164	27000	6	2.8	365	42.6	2.14
Waikaremoana	NZL	-38.77	176.08	138	5400	256	93	2920	3	0.19

706

For Review Only

Table S2. Median correlation coefficients between unstandardized variance components (VC) of chlF, which explain the proportion of variance in algal biomass explained by hourly, daily, and monthly timescales, and water column stability (measured as the absolute temperature difference between surface water temperature and the temperature at a depth of approximately one-tenth of the maximum depth in each lake, such that larger values indicate greater stability) as well as the overall means for high- frequency physical variables in the study. For all physical variables, we calculated the mean of hourly measurements. Bolded values have 95% CIs that exclude zero.

For Review Only

Attribute	Hour VC	Day VC	Month VC
Mean water column stability	-0.03	-0.14	**-0.62**
Mean Temp	0.11	0.16	-0.24
Mean PAR	0.20	0.17	-0.15
Mean Wind	0.09	0.01	**0.45**

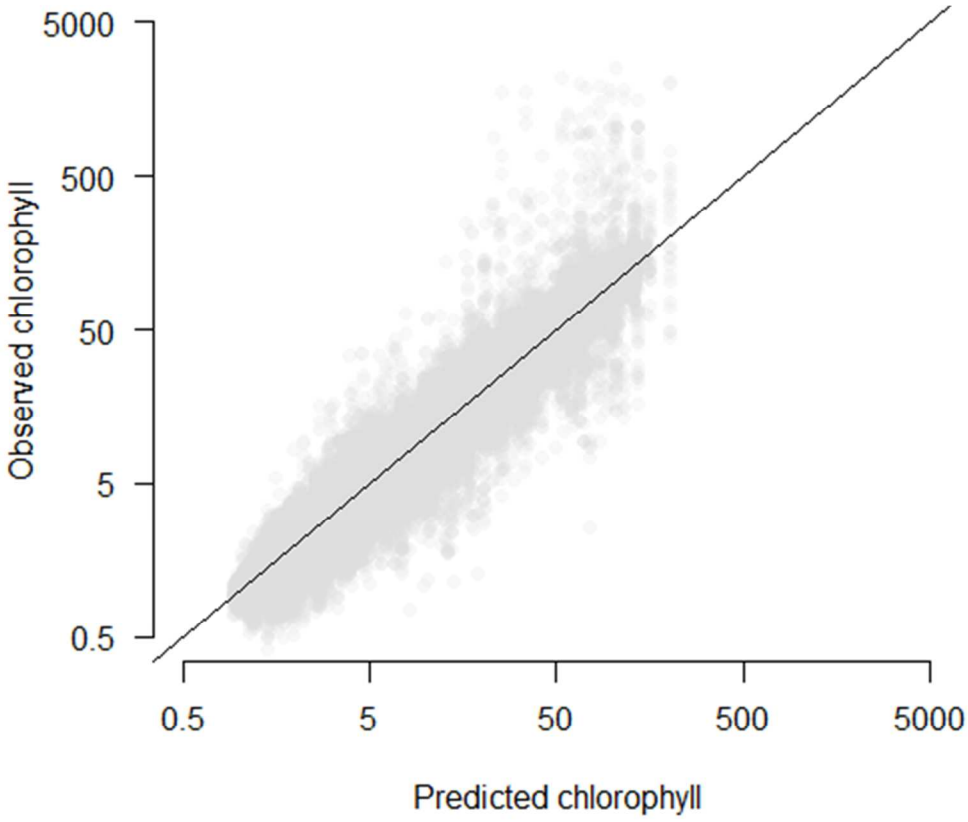
714

716

Figure S1: Global distribution of lakes included in the study.

718

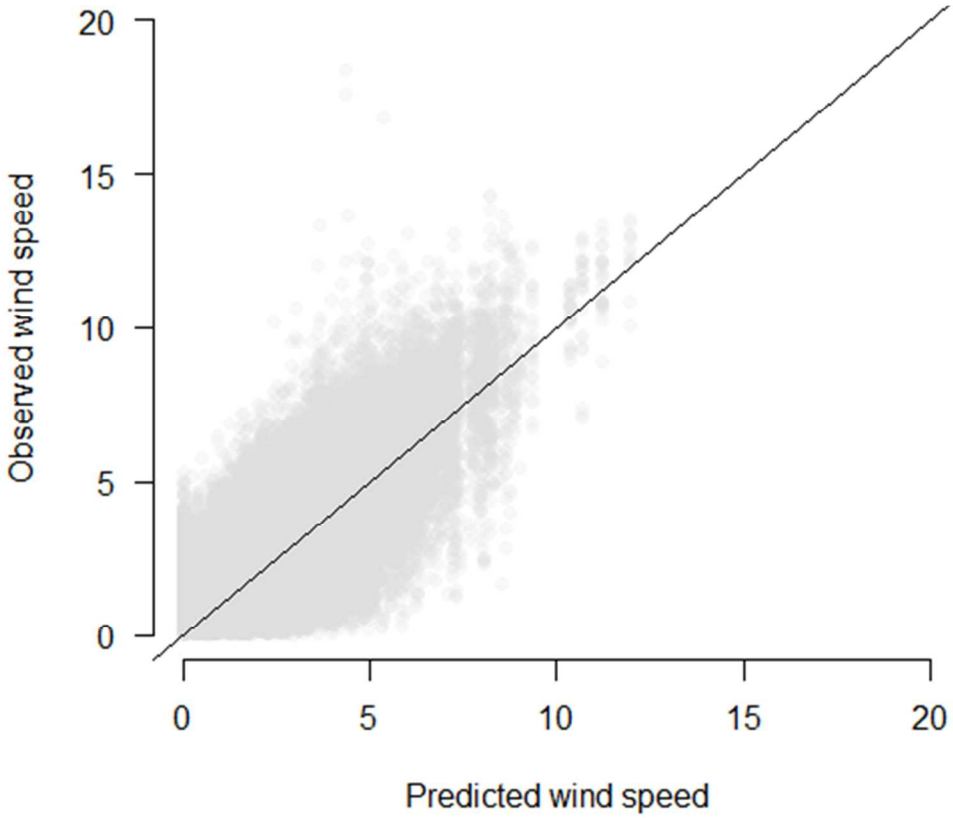
ew Only



720 Figure S2. Model fitted to hourly chlorophyll (µg/L) observations at 18 lakes. Predicted values are posterior means Data are from single years in each lake over the period from 722 2008 – 2013 (n = 94 966). Line is 1:1 relationship, 95% CI for Bayesian R^2 = 0.96 – 0.96

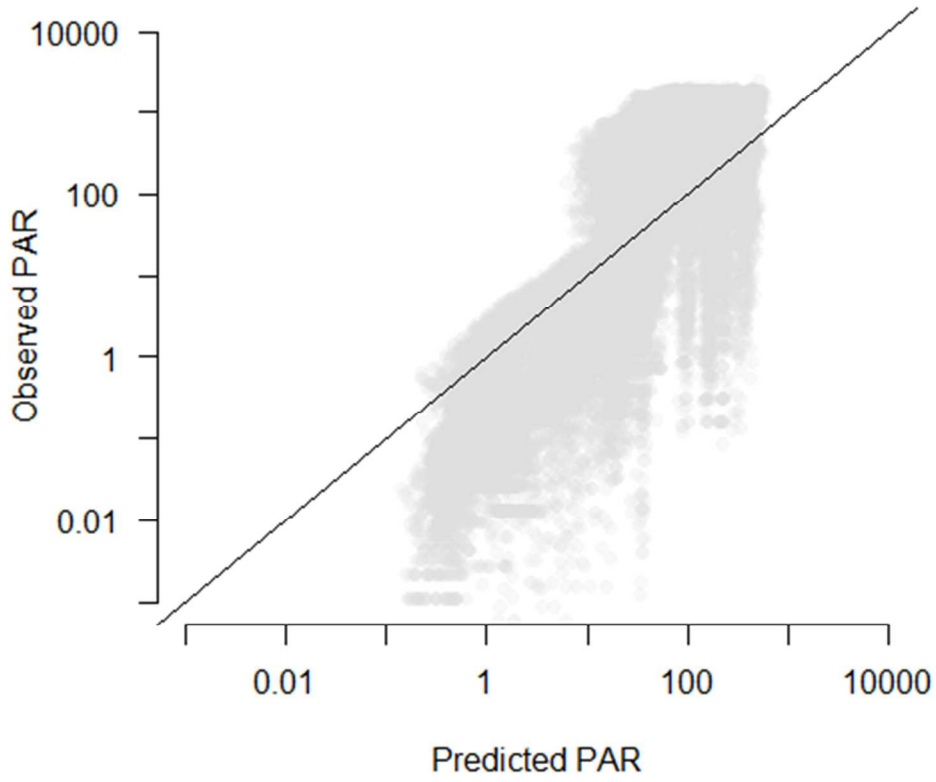
For Review Only

724



726 Figure S3. Model fitted to hourly wind speed (m s^{-1}) observations at 18 lakes (n = 85 649). Line is 1:1 relationship. 95% CI for Bayesian R^2 = 0.44 – 0.48.

For Review Only

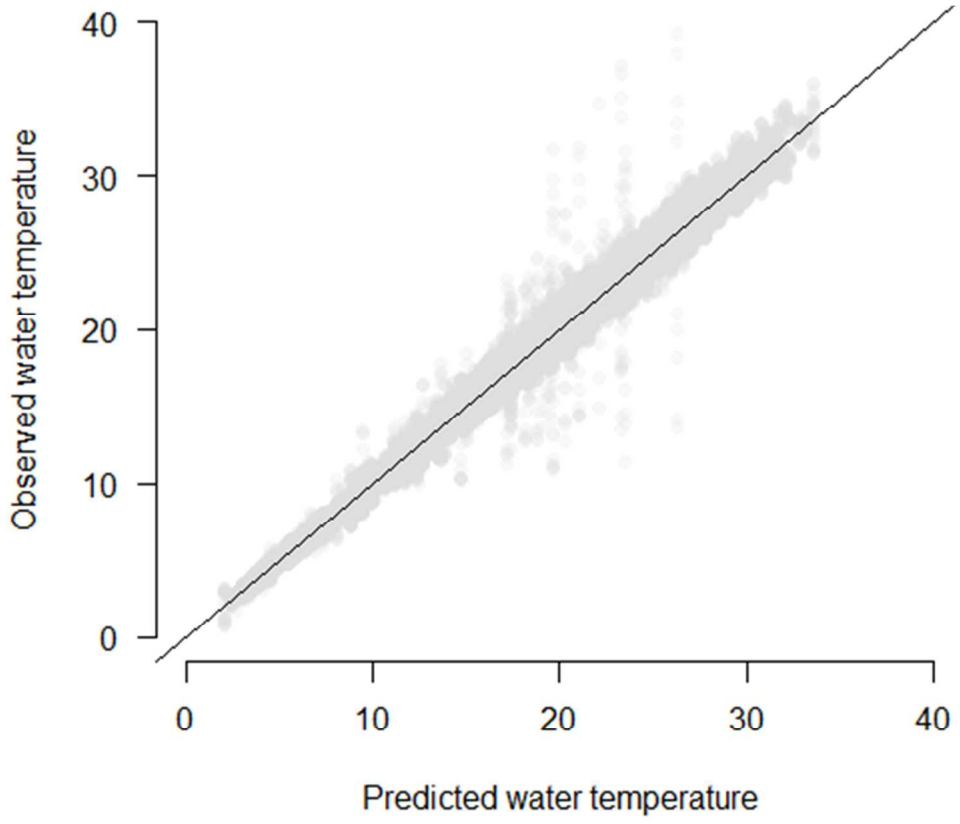


728

Figure S4. Model fitted to hourly PAR observations at 18 lakes (n = 64 540). Line is 1:1

730 relationship, Bayesian R^2 = 0.44 – 0.45.

Review Only



732

Figure S5. Model fitted to hourly surface water temperature (°C) observations at 18 lakes (n = 86

734 538). Line is 1:1 relationship. 95% CI for Bayesian R^2 = 0.994 – 0.994.

For Review Only