Supplemental Material: Annu. Rev. Mar. Sci. 2024. 16.
https://doi.org/10.1146/annurev-marine-041823-095913

Supplemental material

Designing more informative multiple driver experiments

- Mridul K. Thomas & Ravi Ranjan

1. Temperature-nutrient interaction function

For all our simulations and mathematical work in this paper, we used a temperature-nutrient function from Thomas et al. (2017):

$$\mu(T,R) = b_1 \cdot \exp(b_2 \cdot T) \cdot \frac{R}{R+K} - (d_0 + d_1 \cdot \exp(d_2 \cdot T)) \quad (1)$$

The model is fully explained in that paper, and we summarise its explanation here.

Specific growth rate μ depends on temperature, T, and consumable resources (nutrients/food) R. This model brings together the well-known Monod equation for nutrient-dependent growth with a new model for temperature-dependent growth.

The nutrient dependence is captured here by $\frac{R}{R+K}$, where K is the half-saturation constant for growth; note however that K does not have all the same mathematical properties in equation 1 as it does in the basic Monod equation due to its interaction with other terms in the equation. The maximum growth rate of the population is described here by $b_1 \cdot \exp(b_2 \cdot T)$. b_1 is the birth rate at a temperature of 0°C under nutrient-replete conditions and b_2 is the exponential change in birth rate with increasing temperature. The second half of the equation captures mortality rates, where d_0 is a temperature-independent mortality term, while d_1 and d_2 jointly describe the exponential increase in mortality rate with temperature.

This model has empirical support from two phytoplankton studies (Thomas et al. 2017, Bestion et al. 2018), and also captures important variation in food-dependent growth in fish (Brett 1971) and other taxa (Litchman & Thomas 2023). However, not all of its predictions have been carefully evaluated. So as in the case of the temperature-dependent functions described in Box 1, we do not have confidence that this is the true functional form, merely that it appears to

describe variation in growth rate more accurately than previous functions. A different functional form (Huey & Kingsolver 2019) explains the temperature-resource equation in separate terms but leads to a similarly-shaped response surface.

2. Leverage

For any experimental design, an important question is exactly what levels to use for each driver i.e. what treatment combinations are most informative? All data points don't contribute equally to the estimation of regression parameters (Burch et al. 2012; Hoaglin & Welsch 1978; Smith 2005). In a linear regression, an observation farther away from the remaining observations on the X-axis has more *leverage* than the points in the cluster (Chatterjee & Hadi 1986; Cook & Weisberg 1982). If this point is also sufficiently different that it lies away from the regression line that would be drawn in its absence, this far-away point 'pulls' the fitted line strongly towards itself. The strength of this 'pull' of an observation on a fitted curve can be quantified, and is known as an observation's *influence* (Stål 2015). High-leverage points affect the estimated parameter values substantially. Therefore, the leverage and influence of observations is often of interest to experimenters. Note however, that while leverage and influence are informative and can help with the creation of custom designs, optimal designs take into account more information and are superior (see the main text section titled *Optimal Designs*).

For a linear regression, the leverage of a point is calculated via the **hat matrix ($\boldsymbol{H}$)** (Hoaglin & Welsch 1978), which relates the observation vector $\boldsymbol{Y}$ with the fitted values $\widehat{\boldsymbol{Y}}$ (Cook & Weisberg 1982; Emerson et al. 1984; St. Laurent & Cook 1992; Yoshizoe 1991):

$$\widehat{\boldsymbol{Y}} = \boldsymbol{HY} \tag{2}$$

The hat matrix is calculated from the design matrix, $\boldsymbol{X}$ (see Sidebar titled Optimal Design Criteria in the main text):

$$\boldsymbol{H} = \boldsymbol{X}\left(\boldsymbol{X^T X}\right)^{-1}\boldsymbol{X^T} \tag{3}$$

The hat matrix has n rows and p columns. Any element H_{ij} of the hat matrix measures the impact of a change in the j-th observation Y_j on the i-th fitted value $\widehat{Y}_i$. The diagonal elements of this matrix, H_{ii} measure the impact of any observation on the corresponding fitted point, thus measuring the leverage of the observation. For linear regressions, the leverage values are always bounded between zero and one ($0 \leq H_{ii} \leq 1$). A leverage value of zero means that the

fitted value is independent of the corresponding observation. When leverage is one, the fitted value is the same as the observation.

Both the design matrix and the hat matrix depend on the sensitivity ($\partial\eta/\partial\theta_j$) of the function $\eta(\boldsymbol{X};\boldsymbol{\theta})$ to small changes in parameters ($\boldsymbol{\theta}$) (St. Laurent & Cook 1992). In case of a linear regression, the leverage values are independent of the fitted parameters and can be calculated across the entire range of predictor variables before conducting the experiment (Burch et al. 2012). Experimentalists can use leverages to identify the regions of predictor variables where the function is especially sensitive to change in parameters. Directing high sampling effort towards these regions will significantly improve the resulting fit from the experiment.

In a non-linear regression, $\eta(\boldsymbol{X};\boldsymbol{\theta})$ is non-linear with respect to the parameters $\boldsymbol{\theta}$. Therefore, the leverages depend on the fitted parameters. To get around this, researchers have used Taylor approximations of the non-linear function around a reference fitted parameter set ($\widehat{\boldsymbol{\theta}}$) (St. Laurent & Cook 1992). This approximation is known as the tangent plane approximation:

$$y_i \approx \eta\left(x_i;\widehat{\boldsymbol{\theta}}\right) + \frac{\partial\eta\left(x_i;\widehat{\boldsymbol{\theta}}\right)}{\partial\boldsymbol{\theta}}\left(\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}}\right) + \epsilon_i,\ i = 1 \ldots n. \tag{4}$$

This approximation effectively converts the non-linear function to a linear function. The *tangent plane leverage matrix* is then defined as:

$$\boldsymbol{H} = \widehat{\boldsymbol{X}}\left(\widehat{\boldsymbol{X}}^{\boldsymbol{T}}\widehat{\boldsymbol{X}}\right)^{-\mathbf{1}}\widehat{\boldsymbol{X}}^{\boldsymbol{T}} \tag{5}$$

Here, $\widehat{\boldsymbol{X}}$ represents the design matrix $\boldsymbol{X}$ evaluated at the reference parameters $\widehat{\boldsymbol{\theta}}$. The tangent plane leverage matrix is the equivalent of the hat matrix for a linear regression model, so we use the same notation ($\boldsymbol{H}$) for both. As earlier, the diagonal elements of the tangent plane leverage matrix represent the leverage of each observation. The tangent plane leverages also range between zero to one and sum to p, the number of parameters. The tangent plane approximation of nonlinear functions can vary in its accuracy, depending on the nonlinearity of the function involved. When the approximation ceases to be good, the leverages also become less useful.

For nonlinear regressions, tangent plane leverages rely on linear approximations of the nonlinear function. To fully account for the nonlinearity of the functions, researchers have designed alternative methods of measuring leverages, such as the Jacobian leverage (St. Laurent & Cook 1992). The Jacobian leverage matrix ($\boldsymbol{J}$) is given by:

$$\boldsymbol{J} = \widehat{\boldsymbol{X}}\left(\widehat{\boldsymbol{X}}^{\boldsymbol{T}}\widehat{\boldsymbol{X}} - [\hat{\boldsymbol{e}}^{\boldsymbol{T}}][\widehat{\boldsymbol{W}}]\right)^{-1}\widehat{\boldsymbol{X}}^{\boldsymbol{T}} \quad (6)$$

The vector $\hat{\boldsymbol{e}}^{\boldsymbol{T}}$ $(n \times 1)$ represents the residuals associated with the n observations. The matrix $\boldsymbol{W}$ $(p \times p)$ represents the Hessian matrix, which consists of the second order derivatives of the function $\eta(x_i; \boldsymbol{\theta})$ with respect to the parameters $\boldsymbol{\theta}$.

$$W_{ij} = \frac{\partial^2 \eta}{\partial \theta_i \theta_j} \quad (7)$$

$\widehat{\boldsymbol{W}}$ $(n \times p \times p)$ is the tensor resulting from evaluating the Hessian matrix at the fitted parameters $\widehat{\boldsymbol{\theta}}$. The term $[\hat{\boldsymbol{e}}^{\boldsymbol{T}}][\widehat{\boldsymbol{W}}] = \sum_{i=1}^{n} \hat{\boldsymbol{e}}_{\boldsymbol{i}} \widehat{\boldsymbol{W}}_{\boldsymbol{i}}$, $i = 1 \ldots n$ is a tensor contraction with contractions done in the slots one and two following St. Laurent & Cook (1992) and Bates & Watts (1980). As before, the diagonal elements J_{ii} of the Jacobian leverage matrix represent the Jacobian leverage.

The Jacobian leverage matrix takes into account the curvature of the non-linear function via the Hessian matrix $\boldsymbol{W}$, while the tangent plane leverage matrix does not. There are two cases when these two leverages are identical: either if the function is linear ($W_{ij} = 0, i = 1 \ldots p, j = 1 \ldots p$) or if the model fits the data perfectly (${\hat{e}^T}_i = 0, i = 1 \ldots n$). Researchers have explored precisely how the curvature of the function impacts the Jacobian leverage (St. Laurent & Cook 1993).

Since the Jacobian leverage incorporates the curvature of the function being fitted, the leverage values are more accurate for nonlinear models. However, in practice, it can be hard for an experimenter to interpret Jacobian leverage values for several reasons. For example, unlike the tangent plane leverage, Jacobian leverage values can exceed one for a *super-leveraged* observation (St. Laurent & Cook 1992; Stål 2015). Leverage values can be interpreted as the rate of change of the fitted values resulting from a change in the corresponding observed value, if a new and slightly different measurement were to replace the observation. Superleverage occurs when the change in a fitted value is larger than the change in the corresponding observation itself. While interesting theoretically, the values can go up to infinity making it hard to decide what constitutes a 'high' leverage value. The Jacobian leverage values also don't sum up to p, the number of parameters. Perhaps most importantly in the context of experimental design, Jacobian leverage values can *only* be calculated for the predictor variable levels at which the data was collected. Therefore, the Jacobian leverage values give us no information

about where it might be most useful to place experimental points either *a priori* (even with reasonable parameter guesses) or as part of a sequential design.

We therefore recommend the use of the tangent plane approximation for calculating leverage in nonlinear regressions in global change research. Due to their bounded nature, tangent plane leverage values can easily be compared with each other. Further, their values never exceed one, making them easily interpretable and comparable. Finally, given one set of experimental values, the tangent plane allows us to calculate the leverage *even for points not measured in that experiment* (Burch et al. 2012). This can allow the experimenter to design their experiment better by choosing levels likely to be high-leverage *a priori* (if reasonable parameter guesses are possible) or iteratively as part of a sequential measurement-and-fitting process.

While we focus on leverages here, the influence of any observation depends on both its residual and the leverage value. The distinction between leverage and influence is clear in the case of linear regression, where high influence observations can result from having a high leverage predictor value or a high residual or both. However, the terminology is less clear for nonlinear regressions where metrics like the Jacobian leverage incorporate residuals and therefore actually measure influence. As experimenters can only control the experimental design completely, the leverages at different treatment combinations are of the most interest to them. As a result, we focus primarily on the leverages and their use while mostly ignoring the role of residuals. Readers interested in a more thorough treatment of influence metrics in nonlinear regressions should refer to Stål (2015), Wei et al. (1998) and references therein.

Being aware of leverage is helpful in developing intuition and creating better custom designs (see the main text section titled *Custom Designs*). But it is important to note that placing points solely in high-leverage regions will not lead to a good experimental design with low prediction error. A more sophisticated approach to experimental design must take into account how different points *jointly* affect the ability to estimate parameters well or reduce prediction error. This is the domain of *optimal design* (see the main text section titled *Optimal Designs*).

3. Statistical parameter estimation and RMSE estimation process

For each experimental design, we simulated 100 experiments (Figures 2, 4).

For a single simulation experiment, the simulated growth rates had means equal to the "true" values based on the temperature–nutrient response surface (equation 1) with our chosen parameter values, shown below (see supplemental material section titled *Temperature-nutrient interaction function* for a description of the equation).

Parameter	'True' value used in simulations
b_1	0.4
b_2	0.057
d_0	0.68
d_1	0.004
d_2	0.19
K	1

The simulated growth rates also had random 'noise' or error added that is drawn from a normal distribution with a mean of zero and a standard deviation of 0.1.

We then used maximum likelihood to fit the temperature–nutrient equation (equation 1) to the simulated data using the *R* function *mle2* from the *bbmle* package with the *bobyqa* optimiser from the *optimx* package (except in the case of sequential/adaptive designs, where we used a simpler and faster optimisation process). We used a grid-search approach with 3 starting guesses for each of the 6 parameters (and a single guess for the standard deviation estimate), for a total of $3^6 = 729$ sets of starting guesses. To limit the space of possibilities based on reasonable expectations of *a priori* knowledge, we also imposed lower and upper limits using the *bobyqa* optimiser. Starting parameter guesses and the lower and upper limits are shown in the table below.

Parameter	Set of initial guesses used in parameter estimation	Lower limit of parameter search	Upper limit of parameter search
b_1	0.3, 0.6, 0.9	0	5
b_2*	0.04, 0.08, 0.12 (x 10)	0	5
d_0	0.3, 0.6, 0.9	0	5
d_1*	0.002, 0.006, 0.010 (x 100)	0	5
d_2	0.3, 0.6, 0.9	0	5
K	0.5, 1.0, 1.5	0	10
sd †	0.2	0	10

* Note that to avoid algorithmic problems associated with having parameter values on very different scales, we modified equation 1 before fitting by dividing b_2 by 10 and d_1 by 100, and correspondingly multiplied the initial guesses for b_2 by 10 and d_1 by 100. These changes cancel out and are done for algorithmic convenience; they do not change anything meaningful. The values before multiplication are shown here, with the multiplication factor in parentheses.

† Note that the *sd* parameter is simply standard deviation and is not part of the temperature-nutrient equation itself, and so we did not use different starting guesses.

For each simulation, we compared the 729 sets of fits and choose the best one using AIC (essentially just comparing log-likelihood, since all the fits were based on the same equation and so the AIC parameter penalty was irrelevant). In the event of a tie, the winner was chosen randomly.

With the parameter values from the best fit from each simulation, we then calculated the prediction error (root mean square error, RMSE). We first defined an evenly-spaced 100 × 100 grid in temperature-nutrient space, with temperature limits of 10 and 30 °C and nutrient limits of 2 and 10 μM. At each point on the 100 × 100 grid, we compared the predicted growth rate from the best fit parameter values with the true values (from the known true parameter values). We then calculated the RMSE for each simulation. The distribution of these RMSE values across all simulations for all designs is shown in Figure 3.

For most of the designs considered, the exact temperature and nutrient treatment levels were identical in all simulations of the same design. However, these differed between simulations in the space-filling, optimal, and sequential/adaptive designs. In principle the effect of the variation in designs could be investigated further for their impact on RMSE, but this was not our focus and we chose to ignore this here. However, we describe the reasons for the variation in designs

for these three cases here. (i) Space-filling design calculations involve some randomness and so varied with each simulation. (ii) Optimal designs vary based on initial parameter guesses. To account for uncertainty in the parameter guesses, we generated designs based on a range of guesses based on those used in our fitting algorithm (see supplemental material section titled *Optimal design process*). These designs were quite similar but were not identical (Supplemental figure 2). (iii) Sequential/adaptive designs used optimal designs as their starting point and so varied based on initial parameter guesses. Even if different simulations had identical initial designs, however, random noise would lead to slightly different parameter estimates and so subsequent iterations would differ in their point placement (see supplemental material section titled *Sequential/adaptive design process*).

4. Optimal design process

We generated optimal designs with 25 points, for comparison with the similarly-sized full factorial response surface designs. Optimal designs are generated using 'exchange algorithms' that take as their input (i) a specific equation/functional form describing how the response varies based on the predictors, (ii) parameter values (guesses) for all parameters in the function in case of non-linear functions like those we consider in this paper, and (iii) the 'design space' i.e. the set of possible experimental treatment combinations. We address these in turn.

(i) The function we used is specified in equation 1 and described above in supplemental material section titled *Temperature-nutrient interaction function*.

(ii) Because we do not know the parameter values before doing the experiment, we need to generate the optimal design based on parameter guesses. Different guesses lead to different optimal designs. To account for uncertainty in the parameter guesses, we generated designs based on a range of guesses. These guesses were a subset of those we used in our parameter estimation procedure (see supplemental material section titled *Statistical parameter estimation and RMSE estimation process*). From the 729 parameter combinations generated as part of that process, we selected parameter combinations which have maximum growth rates that (a) are positive in at least part of the temperature-nutrient range we focus on (temperature from 10 to 30 °C and nutrients from 2 to10 μM), and (b) do not exceed a maximum bound defined by the Eppley curve, which characterises how the maximum possible growth rate for phytoplankton changes with temperature (Eppley 1972, Bissinger 2008, Kremer et al. 2017). Through this

process, we arrive at 175 parameter sets. We then randomly choose 100 out of the 175 for the optimal design process.

(iii) The design space was chosen with experimental feasibility in mind. We allowed all combinations of temperatures (ranging from 10 and 30 °C with 1 °C intervals) and nutrient concentrations (ranging from 2 to10 μM with 0.2 μM intervals).

With these inputs, we generated I-optimal designs using the KL exchange algorithm (Atkinson et al. 2007). There is no guarantee of a global optimum and as such stopping criteria are arbitrary. Based on our tests, we chose a stopping time of 120 seconds; the best design at this point was our optimal design for that parameter combination. A comparison of the optimal designs for the different parameter guesses shows that despite the large variation in guesses, the optimal designs were very similar (Supplemental figure 2). This robustness makes optimal designs a reasonable option even when there is large uncertainty in the parameter values before the experiment.

5. Sequential/adaptive design process

We generated sequential/adaptive designs with 25 points, for comparison with the similarly-sized full factorial response surface designs and (one-shot) optimal designs.

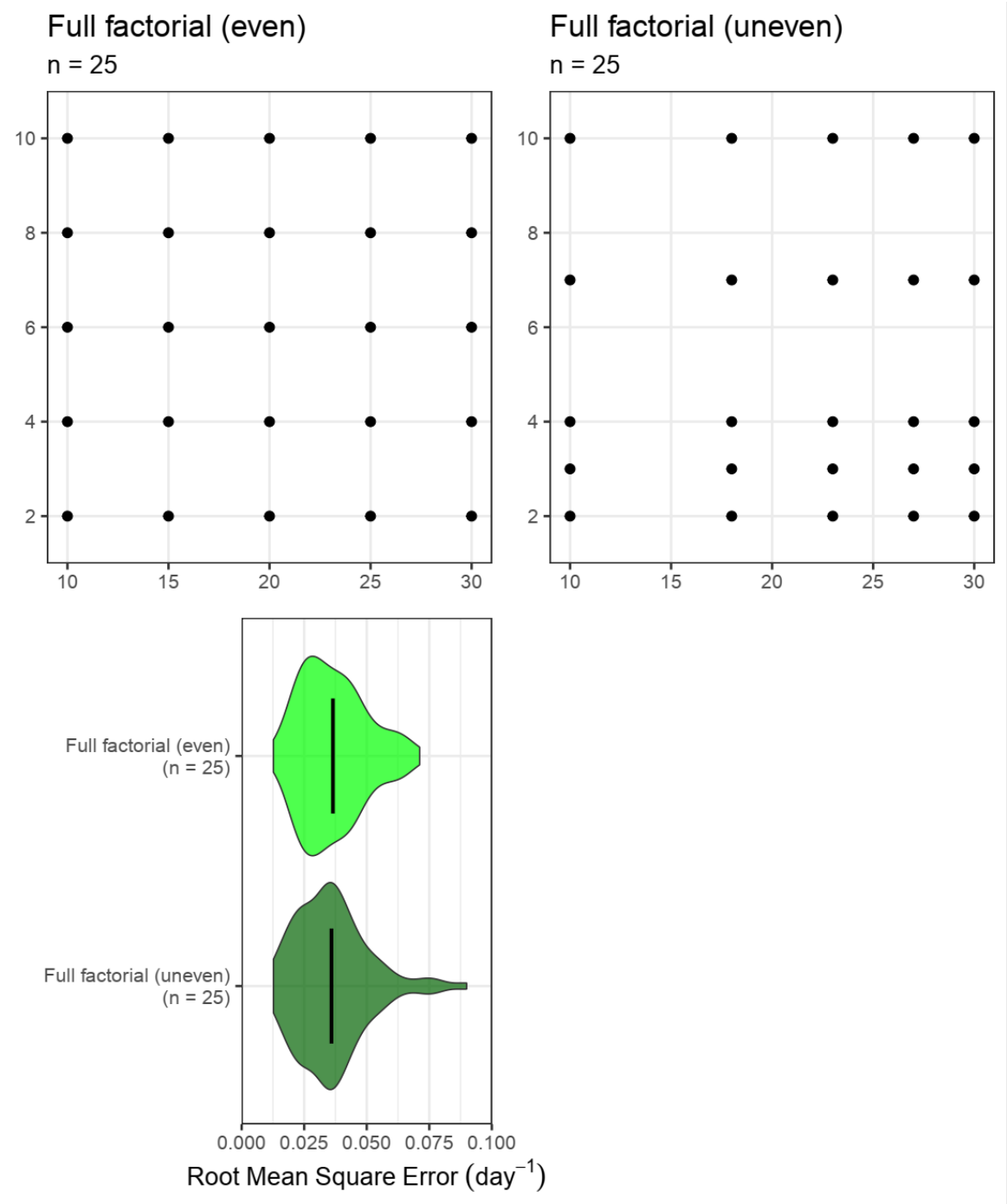
For the first step, we started with an I-optimal experimental design consisting of 15 sampling points. Due to the dependence of the optimal design on the initial parameter guesses, we generate 100 sets of parameter guesses using plausible parameter ranges (the same parameter guess set described in the previous supplemental material section titled *Optimal design process*). We generated the design for the first step using the KL exchange algorithm (Atkinson et al. 2007).

After simulating the growth rates at these 15 points (see supplemental material section titled *Statistical parameter estimation and RMSE estimation process*), we fit the temperature-nutrient function (equation 1) to the data (Figure 4a) and obtained a new set of fitted parameters. For the second step, we modified the KL exchange algorithm so that the sampling points from the previous step remained fixed and the algorithm added five additional sampling points (Figure 4b). After simulating measurements at these five points, we refit the model to the 20 total points and obtained an updated set of parameters. For the third step, we added another five sampling points, taking the total number of points to 25 (Figure 4c). We then again refit the surface to the

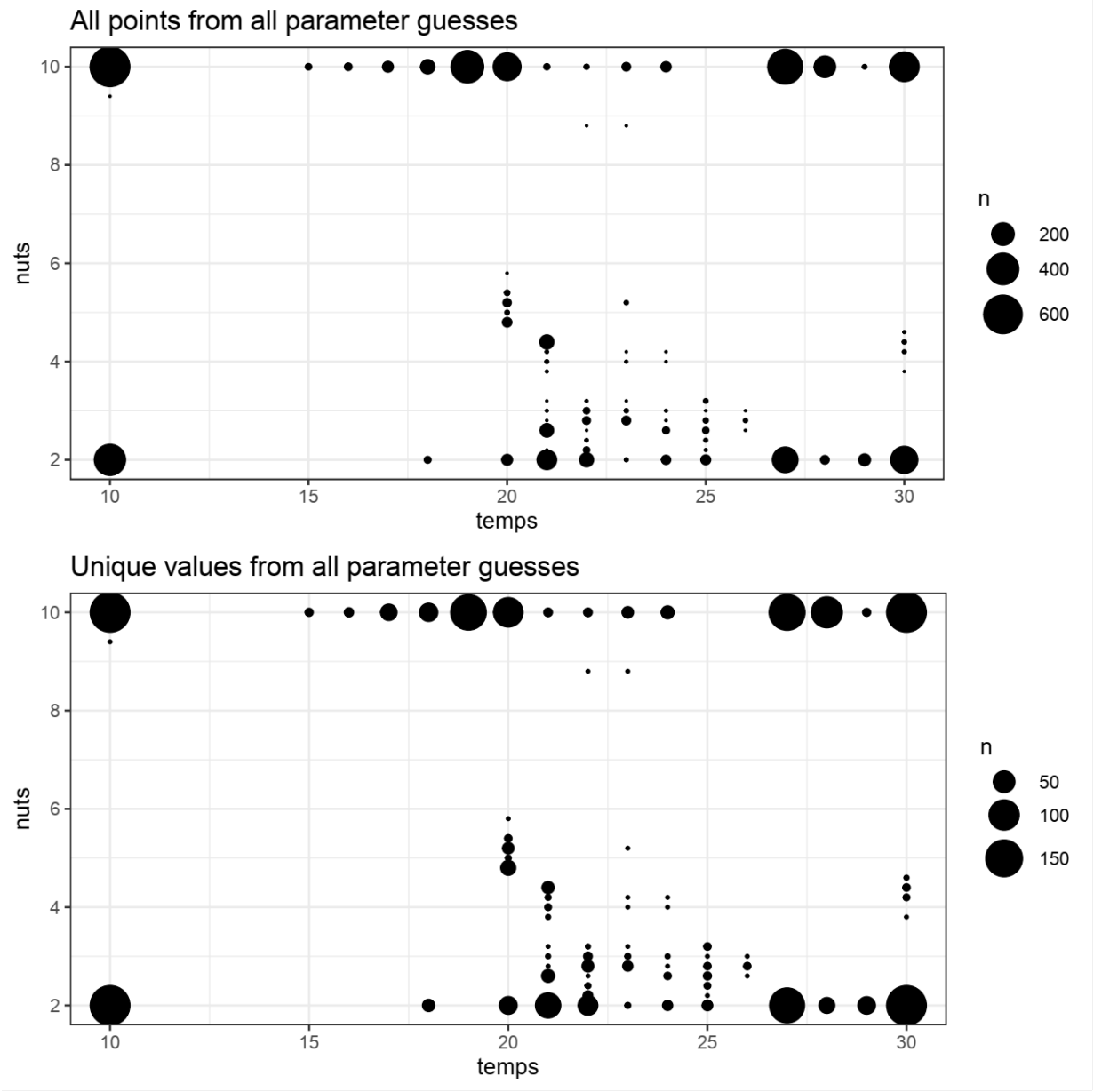
25 total points. We concluded the sequential design process here, but experimenters can choose to add additional points to further improve the fit if desired.

We repeated the entire sequence of steps for each of the 100 initial parameter sets. Within each sequence, we analysed the RMSE each of the three steps (see supplemental material section titled *Statistical parameter estimation and RMSE estimation process*). This resulted in a distribution of RMSEs for each step of the sequential design process. As expected, the mean RMSE decreased with each step in the sequential design process (Figure 4d).

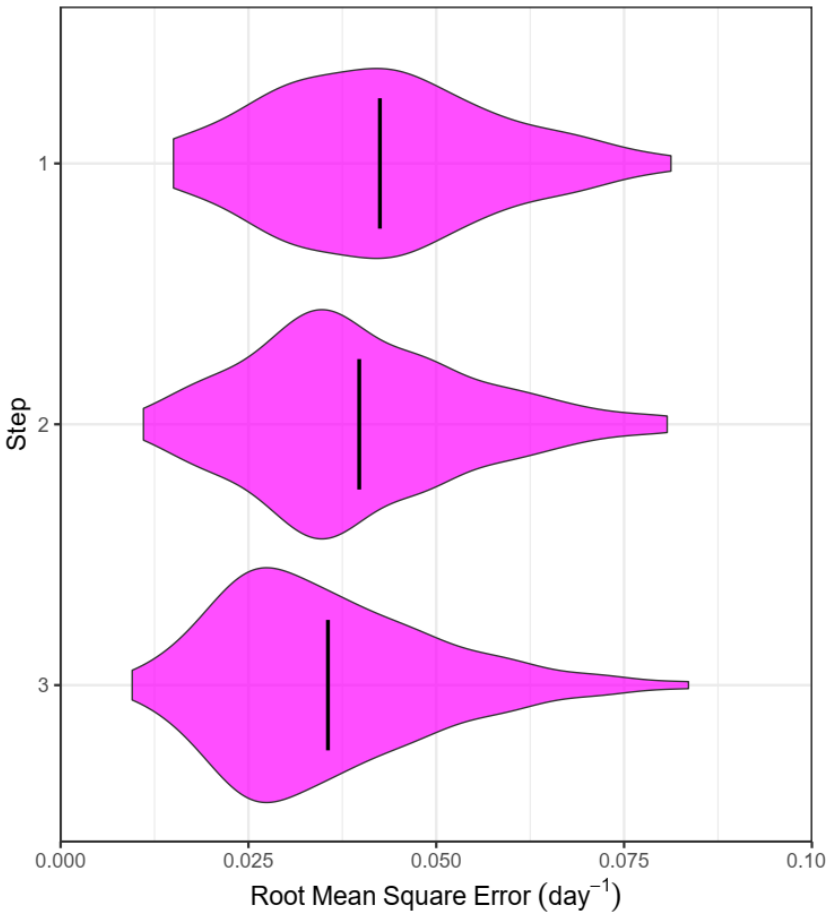
Supplemental figures



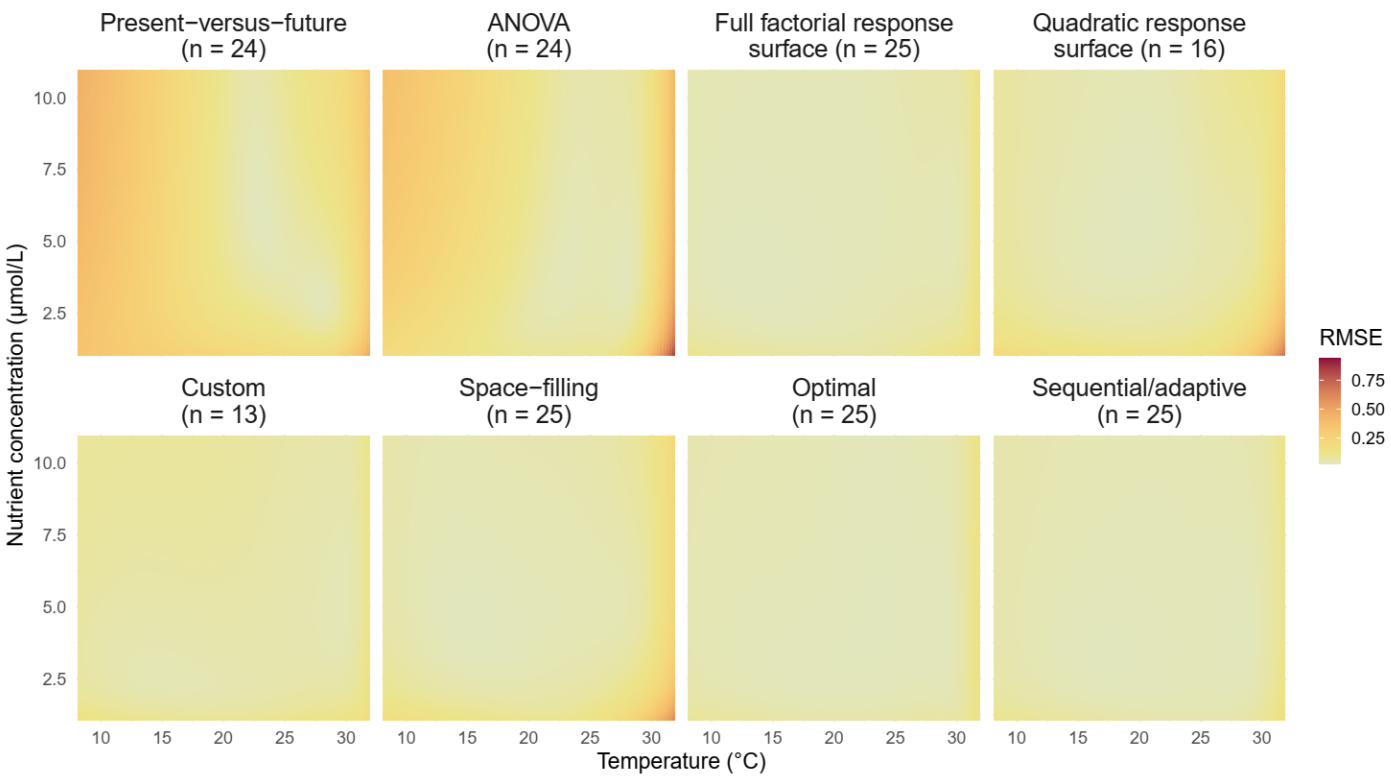
Supplemental Figure 1. A comparison of evenly- and unevenly-spaced full factorial response surface designs shows negligible difference in RMSE. This may differ if the number of levels is smaller, or if the evenly spaced design does not capture the decline in growth rate above the optimum temperature for growth well.



Supplemental Figure 2. Variation in optimal design based on different parameter guesses. These plots show the optimal designs from 175 different plausible parameter combinations, overlaid. The top panel shows all points from the 175 designs. The bottom panel retains only unique values from each design i.e. it eliminates replicates in each design. Overall, the designs show reasonable consistency despite variation in the parameters. The corner points are consistently important, as are other points on the boundary of the design space (the set of possible points, manually defined) at intermediate temperatures. Many of the designs also have one or a few points at low-mid nutrient concentrations between 20 and 27 degrees. These complex but useful designs are unlike anything experimentalists create, and highlight the value of optimal design ideas. We note that this does not imply that an ANOVA design with just the corners would do well, because of uncertainty in the location of the optimum temperature for growth.



Supplemental Figure 3. RMSE decreases with each step of the Sequential/Adaptive Design process. The first step uses 15 points, and 5 points were added for each of the subsequent steps. At each stage, we used an optimal design to choose the positions of the points (see supplemental material section titled *Sequential/adaptive design process* for more details). Note that even with 15 points (step 1), performance is comparable to the 25-point full factorial response surface, indicating how efficient optimal design can be.



Supplemental Figure 4. RMSE variation across the temperature-nutrient surface (slightly extrapolated from the design space on all sides) from all 100 simulations for each design. Lower RMSE (lighter colour) is better. Several designs do poorly at capturing the large decrease in growth rate at high temperature and low nutrient conditions, but optimal, sequential/adaptive, full factorial response surface, and custom designs all achieve this reasonably well. Present-versus-future and ANOVA designs do acceptably over the range over which the points are spread, but extrapolate very poorly despite our constraining the parameters to a realistic range of values.

References

Bates DM, Watts DG. 1980. Relative curvature measures of nonlinearity. *Journal of the Royal Statistical Society. Series B (Methodological)*. 42(1):1–25

Bestion E, Schaum C, Yvon-Durocher G. 2018. Nutrient limitation constrains thermal tolerance in freshwater phytoplankton. *Limnol Oceanogr Lett*. 3(6):436–43

Bissinger JE, Montagnes DJS, harples J, Atkinson D. 2008. Predicting marine phytoplankton maximum growth rates from temperature: Improving on the Eppley curve using quantile regression. *Limnol. Oceanogr.* 53(2):487–93

Brett JR. 1971. Energetic Responses of Salmon to Temperature. A Study of Some Thermal Relations in the Physiology and Freshwater Ecology of Sockeye Salmon (*Oncorhynchus nerka*). *Am Zool*. 11(1):99–113

Burch N, Hoeting JA, Estep D. 2012. Optimal design and directional leverage with applications in differential equation models. *Metrika*. 75(7):895–911

Chatterjee S, Hadi AS. 1986. Influential observations, high leverage points, and outliers in linear regression. *Statistical Science*. 1(3):379–93

Cook RD, Weisberg S. 1982. *Residuals and Influence in Regression*. New York: Chapman and Hall

Emerson JD, Hoaglin DC, Kempthorne PJ. 1984. Leverage in least squares additive-plus-multiplicative fits for two-way tables. *Journal of the American Statistical Association*. 79(386):329–35

Eppley, R. W. (1972). Temperature and phytoplankton growth in the sea. *Fish. Bull.*, 70(4), 1063-1085.

Hoaglin DC, Welsch RE. 1978. The hat matrix in regression and ANOVA. The American Statistician. 32(1):17–22

Huey RB, Kingsolver JG. 2019. Climate Warming, Resource Availability, and the Metabolic Meltdown of Ectotherms. *The American Naturalist*. 194(6):E140–50

Kremer CT, Thomas MK, Litchman E. 2017. Temperature- and size-scaling of phytoplankton population growth rates: Reconciling the Eppley curve and the metabolic theory of ecology. *Limnol. Oceanogr*. 62(4):1658–70

Litchman E, Thomas MK. 2023. Are we underestimating the ecological and evolutionary effects of warming? Interactions with other environmental drivers may increase species vulnerability to high temperatures. *Oikos*. 2023(2):

Smith WF. 2005. *Experimental Design for Formulation*. Society for Industrial and Applied Mathematics

St. Laurent RT, Cook RD. 1992. Leverage and Superleverage in Nonlinear Regression. *Journal of the American Statistical Association*. 87(420):985–90

St. Laurent RT, Cook RD. 1993. Leverage, local influence and curvature in nonlinear regression. *Biometrika*. 80(1):99–106

Stål K. 2015. *Identifying influential observations in nonlinear regression*. Stockholm University

Thomas MK, Aranguren-Gassis M, Kremer CT, Gould MR, Anderson K, et al. 2017. Temperature-nutrient interactions exacerbate sensitivity to warming in phytoplankton. *Global Change Biology*. 23(8):3269–80

Wei B-C, Hu Y-Q, Fung W-K. 1998. Generalized leverage and its applications. *Scandinavian Journal of Statistics*. 25(1):25–37

Yoshizoe Y. 1991. Leverage points in nonlinear regression models. *Journal of Japanese Statistical Society*. 21(1):1–11