What is a good model and how do I get one?

or

Model Calibration and Selection

ESS 211

Steps to Modeling

- Goal Definition
- Conceptualization
- Model Type/Scale Decisions
- Coding
- Calibration ← Data
- Sensitivity Analysis
- "Validation" ← Data

- Application ← Data

Untested Hypothesis

Partially Tested Hypothesis

Calibration vs. Selection

$$Y = F(X, \beta)$$

desired output(s) — functional form of relationship — model input(s) — model parameter(s)

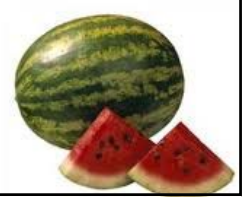
- We will talk about calibration and selection together, since both are aimed at getting a "good" model
- Strictly speaking, calibration is about choosing parameter values (β), selection is more broadly about choosing the structure (F) and parameters.
- For both calibration and selection, the key is being clear about what a "good" model is.

Calibration vs. Selection

$$Y = F(X, \beta)$$

desired output(s) — functional form of relationship — model input(s) — model parameter(s)

- For example, let's consider a case when
  - Y = quality of a piece of fruit (e.g., scale from 1 to 10)
  - X = shape, color, smell, sound
  - F,β = your mental model that relates X to Y



Outline

1) Defining Model Performance
2) Optimizing Model Performance (Calibration)
3) Model Complexity vs. Model Performance
4) Model Assessment and Model Selection
5) Prediction vs. Interpretation
6) An example

Defining Model Performance

What is a good model?

Good Model = Good Prediction = Low Expected Prediction Error
What do we mean by "error"?

Reality: $Y = F(X, \beta) + \varepsilon$

Model: $\hat{Y} = \hat{F}(X, \hat{\beta})$

Loss function: $L(Y, \hat{Y})$

For example: $L(Y, \hat{Y}) = (Y - \hat{Y})^2$ or $L(Y, \hat{Y}) = \text{abs}(Y - \hat{Y})$

Calibration

Population

X, Y

Sampling mechanism

Sample

X_S, Y_S

Future Values

$\tilde{X}$

calibration

$\hat{F}$

prediction

$\hat{Y}$

Define loss function

Calibration

Key points:

1. Structure of F is pre-defined
2. Parameters β depend on sample
3. Goal is to predict well on future sample, not current sample

Key assumptions:

1. **Representativeness**
   -Xs, Ys are good representations of population

2. **Stationarity**
   - relationship between X and Y doesn't change in time

Population

X, Y

Sampling mechanism

Sample

X_S, Y_S

Future Values

$\tilde{X}$

calibration

$\hat{F}$

prediction

$\hat{Y}$

So as one approach we could try to adjust β in order to minimize a loss function on the data

$$L(Y, \hat{Y}) = (Y - \hat{Y})^2 \quad \text{or} \quad L(Y, \hat{Y}) = \text{abs}(Y - \hat{Y})$$

Least-squares means we minimize the first loss function.

e.g. for a linear regression model...

for a linear regression model...

Slope (β_1) = $\frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x - \bar{x})^2}$

Intercept (β_0) = $\bar{y} - \beta_1 \bar{x}$

In-class Exercise!:

Confirm for yourself that this gives the same result as using the built-in linear regression function in R (lm)…

--write function to return slope and intercept from x and y

--generate:

x = rnorm(100); y= 3*x + rnorm(100)

--call your_function(x,y) and compare to lm(y~x)

Slope (β_1) = $\frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sum(x-\bar{x})^2}$

Intercept (β_0) = $\bar{y} - \beta_1\bar{x}$

For linear regression, it is possible to find the minimum least square solution analytically, which is a main reason it is so popular

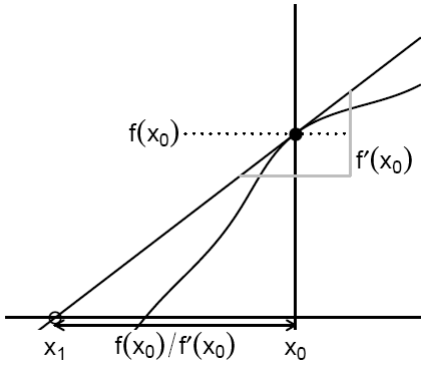
But it is also possible to find solutions numerically. A wide range of algorithms exist to do this.

-brute force (try all possible parameter combinations – works well only if small (<5) # parameters)

-gradient search (or hot-cold) methods (e.g. Newton's)

-others

Newton's Method



Newton's method to find the root of a function.
Note: If we want to minimize a loss function, we can apply Newton's method to the derivative of that loss function.

An example of calibration…

(which you'll use in your HW)

Why not just stop at Least Squares?

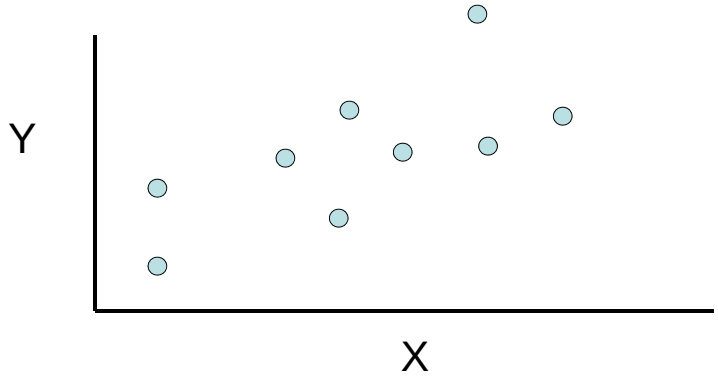
1. Sometimes it is hard to interpret the model and we want to identify a subset of variables/parameters that are most important

2. Sometimes the prediction error is too high, and can be lower for values of β that don't minimize least squares.

How is #2 possible?

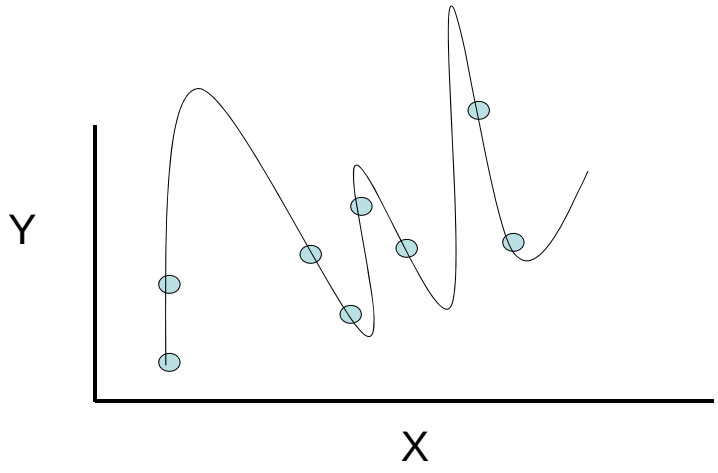
Calibration

There are an infinite number of possible f(X) that would fit the data perfectly.



Calibration

Is this a good model? Why or why not?



Bias-Variance Tradeoff

$$E[(Y - \hat{Y})^2] = E[(F(X) + \varepsilon - \hat{F}(X))^2] = \ldots$$

... = σ_ε^2 + bias^2 + variance

Bias = average error of model over all possible training samples
Variance = variance of error of model over all possible training samples

Models calibrated with least squares have no bias, but can be very fickle (high variance)

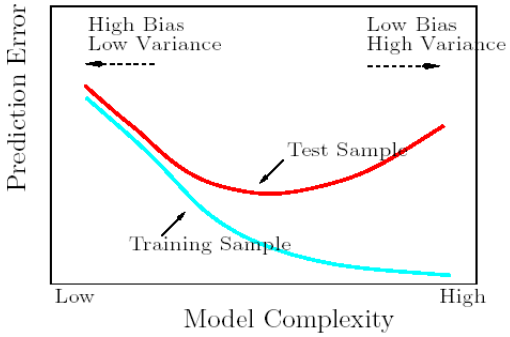


Figure 7.1: *Behavior of test sample and training sample error as the model complexity is varied.*

Hastie et al. 2001

How to avoid high model variance (i.e. overfitting)?

Use low number of parameters (k) relative to sample sizes (n)

Use calibration methods that are less prone to overfitting than least squares

Alternatives to Least Squares:

Maximum Likelihood Estimates (MLE)

i.e. choose the parameters β that make the data most likely

These require an assumption about the distribution of errors, since you need to specify the probability of a data (D) given a hypothesis (H)

Alternatives to Least Squares:

Bayesian Methods

2 unique features of Bayesian methods:

-parameters are treated as distributions, not single values

-able to incorporate prior guesses on parameters – i.e. the result is partly influenced by the data, partly by the prior

Review so far...

- Good Prediction = Low Expected Prediction Error
  = small $E[L(Y,\hat{Y}(\beta,X))]$
- Calibration = Using Sample Data to Adjust Parameters (β)
- Calibration Error ≠ Prediction Error
- E[Prediction Error] = $(\text{Model Bias})^2$ + Model Variance + σ^2_ε

- Some models (and techniques that we use to calibrate models) have low bias but high variance

- Sometimes adding a little bias into a model or calibration method can help reduce total error, by not allowing the model to be too sensitive to the particular sample (i.e. lower model variance)

Outline

1) Defining Model Performance
2) Optimizing Model Performance (Calibration)
3) Model Complexity vs. Model Performance
4) Model Assessment and Model Selection
5) Prediction vs. Interpretation
6) An example

Model Assessment and Selection

-Traditionally, it is common to try to pick a single model to move forward in the process. How does one do this? (next...)

-It is better, though, to keep thinking about multiple hypotheses, and even making predictions with multiple models (we'll return to this when we discuss predictions)

Model Assessment and Selection

-Let's suppose you have N calibrated models

-How would you select among N different models?

-One common one is to pick the one that is expected to give the most accurate predictions. How do we measure this?

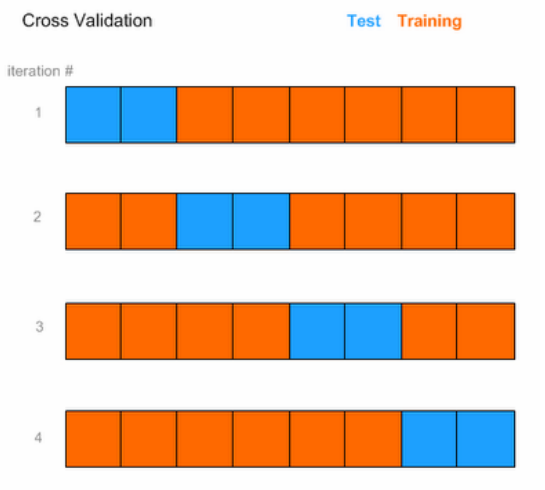
-How about just the fit to sample data (e.g. R^2 or rmse)?

How to assess model prediction error?

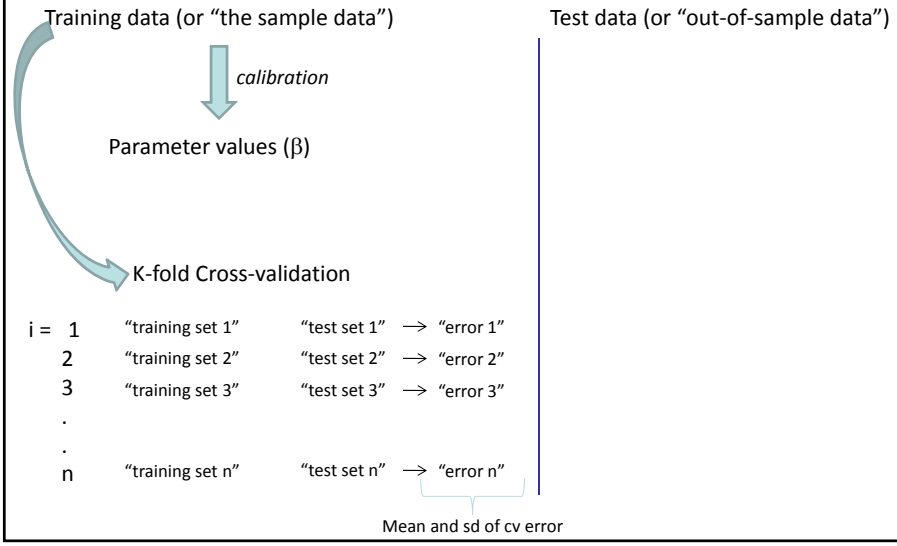
- Split data into training set (~70%) and testing set (~30%)
  - -best option if data is plentiful
  - -need to ensure training and testing data are independent

- Cross-validation (still use all data for calibration, but estimate prediction error by splitting repeatedly)
  - -leave k data points out of calibration
  - -calibrate on remaining n-k points
  - -make predictions for k points
  - -compute error
  - -repeat for different subsets of k
  - -look at average and variance of cross-validated error

Cross-validation: a way to assess prediction error

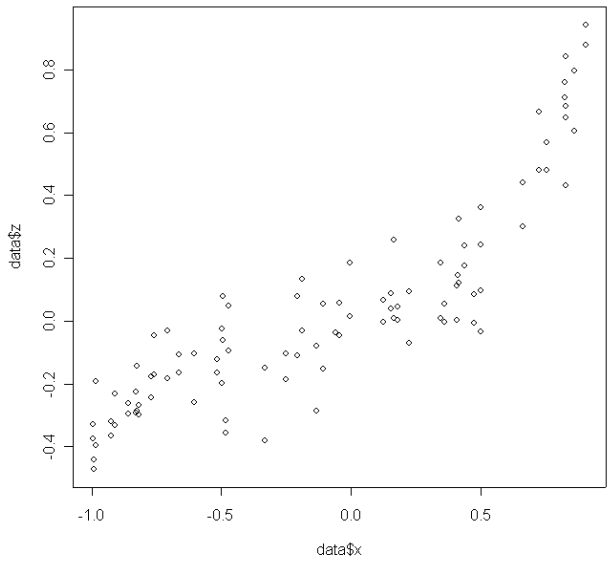


•Cross-validation is trying to mimic the presence of test data without actually having it.
•Don't get confused by reference to "training" and "test" data in cross-validation

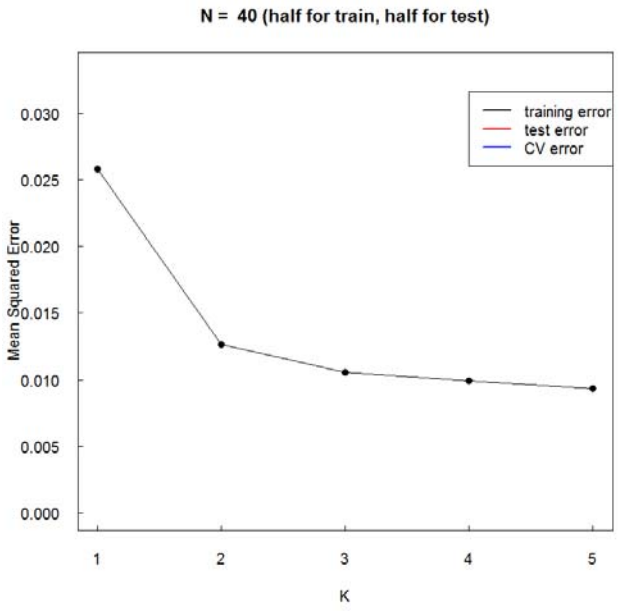


An example of model complexity vs. model performance

- $y = a*x + b*x^2 + c*x^3 + \text{random noise}$
- a=.2, b=.3, c=.5
- observe 100 values of x and y for (-1 < x <1)
- split data into training and test set

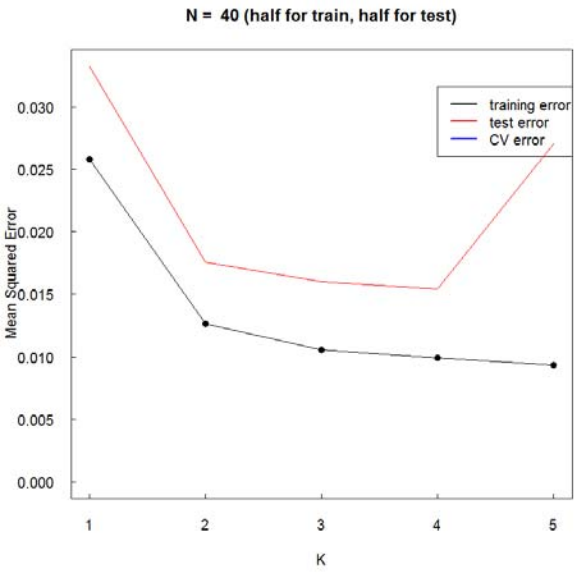


What do you expect to happen if we fit a 1st, 2nd, ... or 5th order polynomial to data?



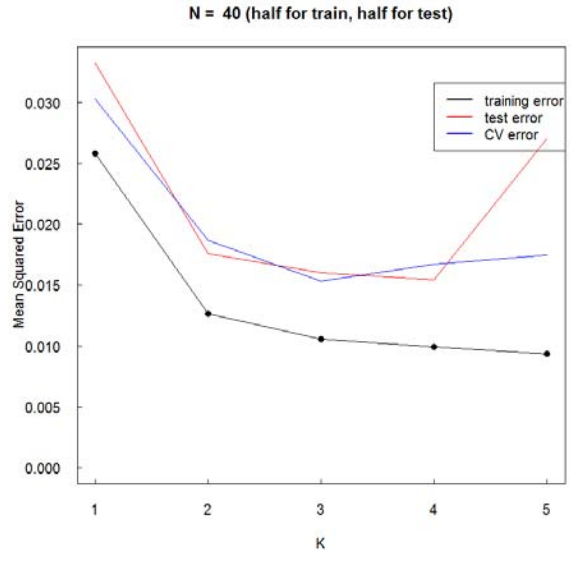
N-fold CV (leave-one-out CV)

Order of polynomial fit to data



N-fold CV (leave-one-out CV)

Order of polynomial fit to data



N-fold CV (leave-one-out CV)

Order of polynomial fit to data

Some other approaches to selection

-Cross-validation is very useful.

-But there are less computationally demanding rules of thumb that are used to avoid overfitting when choosing number of parameters:

-Adjusted $R^2 = 1 - (1-R^2)*(n-1)/(n-p-1)$,

p = # parameters, n = # observations

-Akaike Information Criterion = 2*p -2ln(L)

Summary

- E[prediction error] = σ_ε^2 + bias^2 + variance
- Model variance refers to how much the model changes based on the particular sample (it doesn't mean the variance of the model predictions). It is not bad per se, but shouldn't be too high.
- Model variance is influenced by
  - # of observations used in calibration
  - Complexity of model (# of parameters)
  - If and how priors are used in calibration
- K-fold cross-validation can provide useful estimate of E[prediction error]. In general, k should not be too small or big.
- Also (we haven't discussed much): If a model has many parameters, it is very likely that different combinations of parameters can give equally good calibration errors. (more later when we discuss model evaluation)

So let's test our understanding of bias-variance tradeoffs and cross-validation:

We expect that techniques that add bias can help to avoid overfitting, or high model variance, but they add bias...

What happens if we repeat previous example but we add bias using ridge regression, starting with a little bias and moving to a lot...

An example: calibrating streamflow models

ELSEVIER

Journal of Hydrology 242 (2001) 275–301

Journal of Hydrology

www.elsevier.com/locate/jhydrol

Does a large number of parameters enhance model performance? Comparative assessment of common catchment model structures on 429 catchments

C. Perrin*, C. Michel, V. Andréassian

Water Quality and Hydrology Research Unit, Cemagref, Parc de Tourvoie BP 44, 92163 Antony Cedex, France

Received 21 February 2000; revised 10 July 2000; accepted 30 October 2000

Table 1
Ranges in size and hydro-climatic characteristics for the sample of 429

Country	Australia	Brazil	France	Ivory coast	United States
Number of catchments	26	4	307	10	82
Catchment area (km^2)	3–2500	11300–50600	1–43800	207–6830	0.1–9890
Mean annual potential evapotrasnpiration (mm)	790–1850	930–1080	630–1250	1330–1770	680–2040
Mean annual precipitation (mm)	300–2100	1450–1580	570–2300	1060–1630	300–1540
Mean annual streamflow (mm)	2–1460	370–590	26–2040	32–460	0.2–840
Catchment yield (%)	0.3–71.3	23.1–38.1	3.7–153	2.6–26.5	0.1–72.6
BFI (%)	0.1–82.3	62.3–79.1	5–98.5	34–74.9	0.1–81.2
Irregularity coefficient of rainfall (%)	158–497	284–290	144–329	235–313	146–447
Irregularity coefficient of streamflow (%)	222–1170	208–260	39–673	173–504	147–1110

An example: calibrating streamflow models

Table 2
List of model structures with the retained number of parameters

Original model name and/or reference	Number of optimized parameters in the tested version
Tsykin (1985)	5
GR3J (Edijatno et al., 1999)	3
Model 16 (Bonvoisin and Boorman, 1992)	5
Model 15 (Bonvoisin and Boorman, 1992)	6
PDM (Moore and Clarke, 1981)	6
IHACRES (Jakeman et al., 1990)	7
TANK (Sugawara, 1995)	7
TOPMODEL (Beven and Kirkby, 1979)	7
MODGLO (Servat, 1986)	8
mSFB (Summer et al., 1997)	8
SMAR (Tan and O'Connor, 1996)	8
Wageningen (Warmerdam et al., 1997)	8
Xinanjiang (Zhao and Liu, 1995)	8
Arno (Todini, 1996)	9
Dawdy and O'Donnell (1965)	9
Georgakakos and Baumer (1996)	9
HBV (Bergström, 1995)	9
Institute of Hydrology lumped model (Blackie and Eeles, 1985)	9
MODHYDROLOG (Chiew and McMahon, 1994)	9
NAM (DHI, 1996)	9

An example: calibrating streamflow models

For each catchment, the models were successively calibrated on each sub-period and then tested in verification mode on all the remaining periods. For example, on a catchment with six test periods, six calibration and 30 verification tests were performed. This represents a total of 3204 verification tests for the 429 catchments. To our knowledge, this is the most

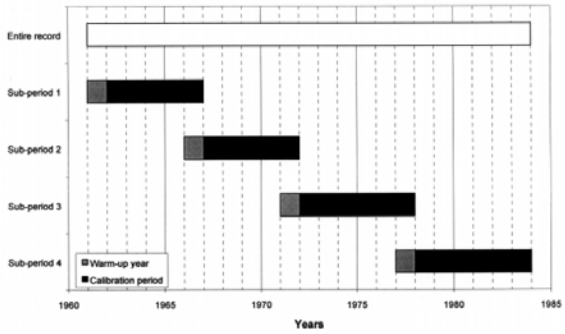


Fig. 4. Example of data record splitting into two sub-periods of six years and two sub-periods of seven years including a warm-up year.

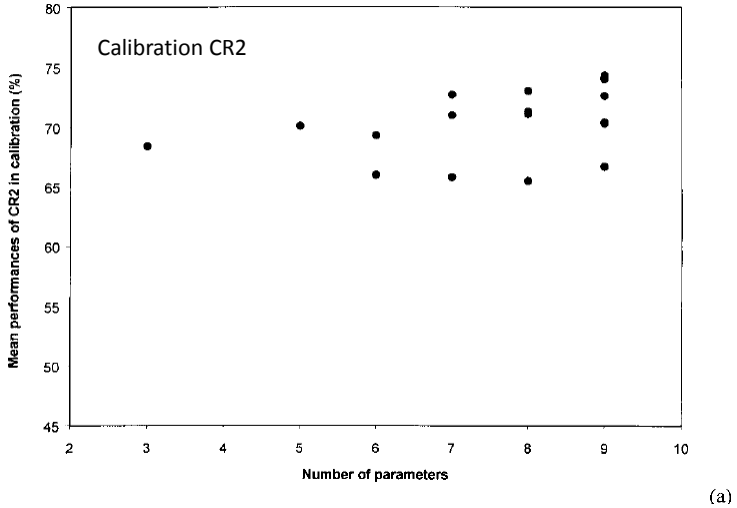
An example: calibrating streamflow models

$$\mathrm{CR2}(\%) = 100\left(1 - \frac{\sum_{i=1}^{n}(\sqrt{Q_{\mathrm{obs},i}} - \sqrt{Q_{\mathrm{cal},i}})^2}{\sum_{i=1}^{n}(\sqrt{Q_{\mathrm{obs},i}} - \overline{\sqrt{Q_{\mathrm{obs}}}})^2}\right) \quad (5)$$

This is the criterion used here to calibrate all models. Hence, the model structures are required

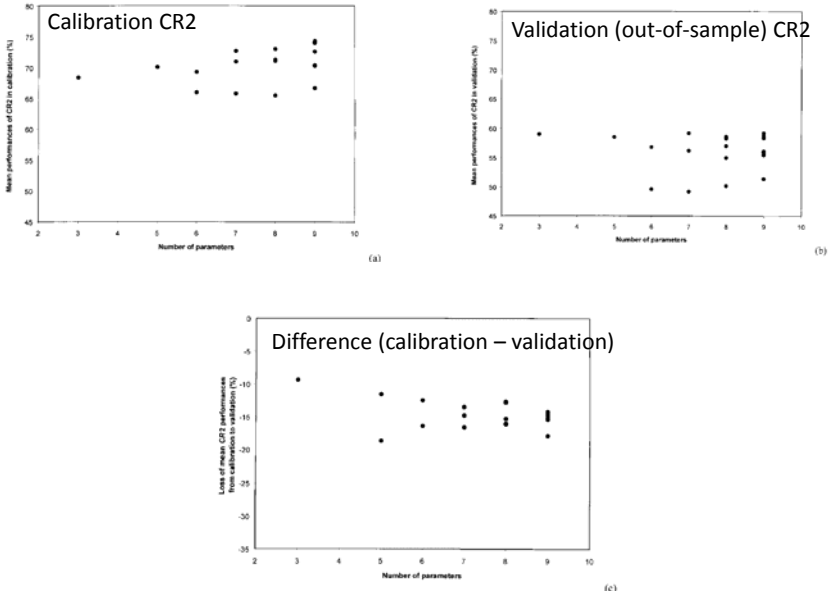
The selected optimisation technique is the steepest descent method summarised in Edijatno et al. (1999). In this method each optimisation run starts with an initial parameter set identified for each model as the one yielding the best results on the whole sample of catchments. Then the algorithm evolves step-by-step in the parameter space toward the 'optimum' parameter values. Outside the scope of this paper, the application of the algorithm to four model structures with different numbers of parameters, showed that the combined use in calibration of two different initial parameter sets for each structure did not produce significant improvements in model performances in verification mode.

An example: calibrating streamflow models



(a)

An example: calibrating streamflow models



An example: calibrating parameter as point values vs. distributions

-Modeling rice yield response in Asia to climate change

Part A: Using the process-based model Oryza

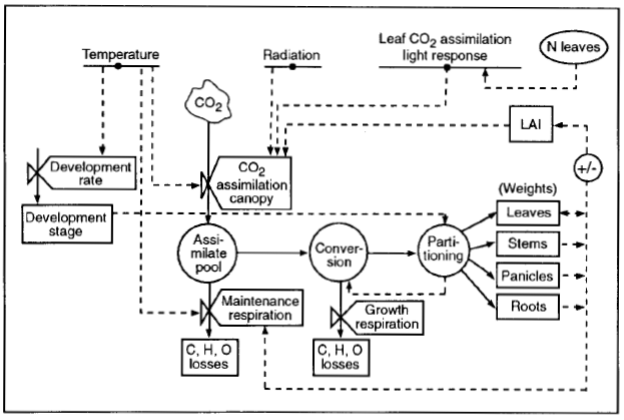


Fig. 3.1. A schematic representation of the model ORYZA1. Boxes are state variables, valves are rate variables, and circles are intermediate variables. Solid lines are flows of material and dashed lines are flows of information.

An example:

Table 7.1. Genotype coefficients and function tables (expressed as a function of developmental stage, DVS) for the two varieties, IR64 and Ishikari, used in the ORYZA1 simulations. FSHTB, FLVTB, and FSTTB are the fractions of new assimilate partitioned to the shoot, leaves, and stem respectively; SLATB is the specific leaf area (cm g^{-1}), DRLVT is the relative death rate of the leaves (g g^{-1}), and NFLVTB is the leaf nitrogen concentration (g N m^{-2} leaf).

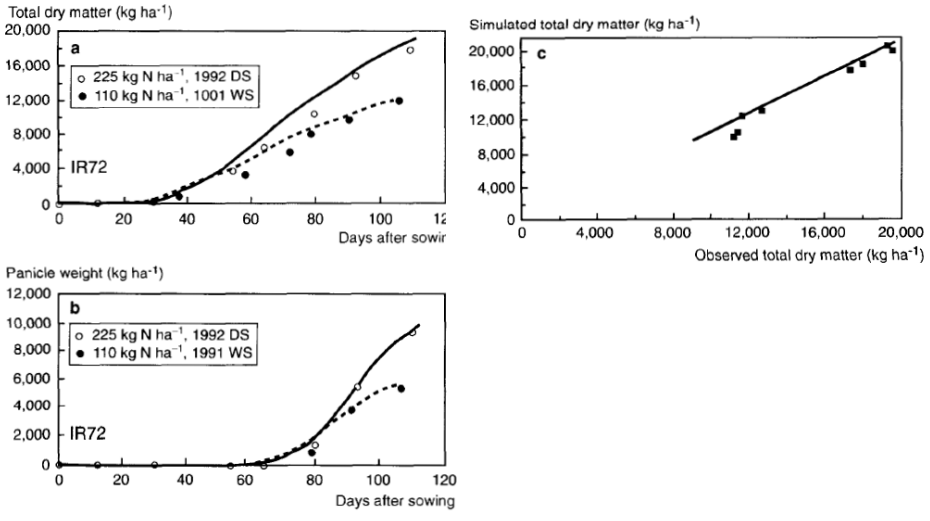
Genotype coefficient	Variable name	Units	Genotype	
			IR64	Ishikari
Initial relative leaf area growth rate	RGRL	$(°Cd)^{-1}$	0.00901	0.00901
Stem reserves fraction	FSTR	-	0.4	0.4
Basic vegetative period duration	JUDD	Dd	19.5	9.0
Photoperiod sensitive phase duration	PIDD	Dd	15.0	15.0
Minimum optimum photoperiod	MOPP	h	11.7	-
Photoperiod sensitivity	PPSE	$Dd\,h^{-1}$	0.17	0.00
Duration of panicle formation phase	REDD	Dd	25.0	20.0
Duration of grain-filling phase	GFDD	Dd	23.0	33.9
Maximum grain weight	WGRMX	mg $grain^{-1}$	22.8	22.8
Spikelet growth factor	SPGF	sp g^{-1}	49.2	49.2

FSHTB = 0.0, 0.5, 0.43, 0.75, 1.0, 1.0, 2.1, 1.0
FLVTB = 0.0, 0.56, 0.2, 0.56, 0.7, 0.5, 0.9, 0.3, 1.0, 0, 2, 0
FSTTB = 0.0, 0.44, 0.2, 0.44, 0.7, 0.5, 0.9, 0.5, 1.2, 0, 2, 0
SLATB = 0.000, 470, 0.152, 470, 0.336, 330, 0.653, 280, 0.787, 210, 1.011, 190, 1.431, 170, 2.011, 170
DRLVT = 0, 0, 0.6, 0, 1, 0.015, 1.6, 0.025, 2.1, 0.05
NFLVTB = 0.000, 0.200, 0.157, 0.542, 0.333, 1,530, 0.650, 1.221, 0.787, 1.556, 1.000, 1.288, 1.458, 1.373, 2.000, 0.836, 2.100, 0.737

Matthews et al. 1995

An example:

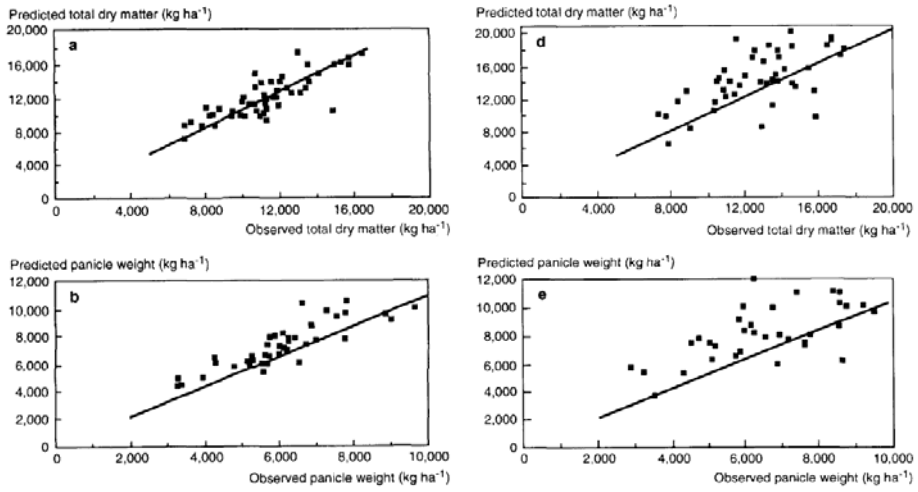
Oryza evaluation for experiments at IRRI



Matthews et al. 1995

An example:

Oryza evaluation for 2 main varieties from unpublished experimental data



Matthews et al. 1995

An example:

Table 7.9. Estimated changes in total rice production predicted by the ORYZA1 model for each country and in the region under the three GCM scenarios. Current actual production ('000 t) in each AEZ on a country basis are adjusted by the simulated changes in total annual production. See text and Matthews *et al.* (1995b) for explanation.

		Current[1]	GFDL		GISS		UKMO	
Country	AEZ	'000 t	% change	'000 t	% change	'000 t	% change	'000 t
Bangladesh	3	27,691	14.2	31,621	−5.0	26,298	−2.8	26,919
China	5	8,854	−7.4	8,201	0.3	8,881	−25.2	6,619
	6	79,872	0.8	80,484	−21.7	62,514	−19.5	64,334
	7	91,828	5.8	97,196	5.8	97,135	3.1	94,695
	8	2,361	−6.4	2,209	−14.2	2,026	−27.6	1,710
India	1	32,807	4.6	34,305	−10.8	29,272	−5.5	31,017
	2	49,949	1.8	50,849	−2.9	48,493	−7.9	46,002
	5	227	−7.4	210	0.3	228	−25.2	170
	6	26,628	5.4	28,069	3.2	27,480	−1.3	26,287
	8	1,011	−6.4	946	−14.2	867	−27.6	732
Indonesia	3	44,726	23.3	55,155	9.0	48,748	5.9	47,387
Japan	8	12,005	−6.4	11,231	−14.2	10,300	−27.6	8,696
Malaysia	3	1,744	24.6	2,173	17.6	2,050	26.8	2,211
Myanmar	2	13,807	21.5	16,776	−10.5	12,356	1.2	13,974
Philippines	3	9,459	14.1	10,797	−11.8	8,340	−4.7	9,018
South Korea	6	8,192	−13.6	7,078	−5.3	7,755	−21.9	6,401
Taiwan	7	2,798	11.8	3,128	12.8	3,156	28.0	3,583
Thailand	2	20,177	9.3	22,044	−4.7	19,230	−0.9	19,989
Total		434,136		462,472		415,129		409,743
% change				6.5		−4.4		−5.6

[1] Source: IRRI (1993).

Matthews et al. 1995

An example:

-Modeling rice yield response in Asia to climate change

Part B: Calibrating a distribution of parameter values (Bayesian) rather than single values (Iizumi et al. 2009)

-Start with a generic model of crop growth with P parameters

Table 2 – Estimated parameters of the large-scale crop model for paddy rice

Abbreviation	Definition	Unit	Likelihood
DVI_0	Initial developmental index (DVI)	Day^{-1}	π_H
G	Minimum number of days required for heading under 350 ppm of atmospheric CO_2 concentration	Day	π_H
A_T	Sensitivity of developmental rate (DVR) to air temperature	–	π_H
T_h	Air temperature at which DVR is half of the maximum rate at the optimum temperature	°C	π_H
B_L	Sensitivity of DVR to day length	–	π_H
L_c	Critical day length	h	π_Y
DVI^*	Value of DVI at which point the crop becomes sensitive to the photoperiod	Day^{-1}	π_Y
LAI_0	Initial leaf area index	–	π_Y
DW_0	Initial dry weight	$g\ m^{-2}$	π_Y
T^*	Base air temperature for calculating cooling degree days	°C	π_Y
C_{cool}	Curvature factor of spikelet sterility caused by low temperature	–	π_Y
C_{hot}	Curvature factor of spikelet sterility caused by high temperature	–	π_Y
r	Technical coefficient	–	π_Y

Note: π_H and π_Y indicate the likelihood functions of heading day and yield, respectively (see Section 4.2).

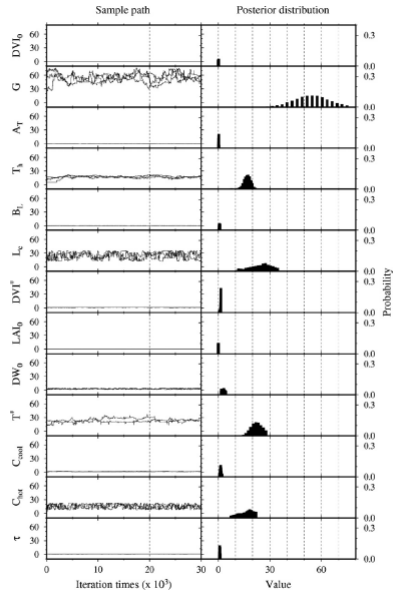


Fig. 3 – Sample paths and posterior distributions of parameters for Aomori provided by the Metropolis–Hastings algorithm. The dashed line in the left panel indicates a burn-in of 20,000.

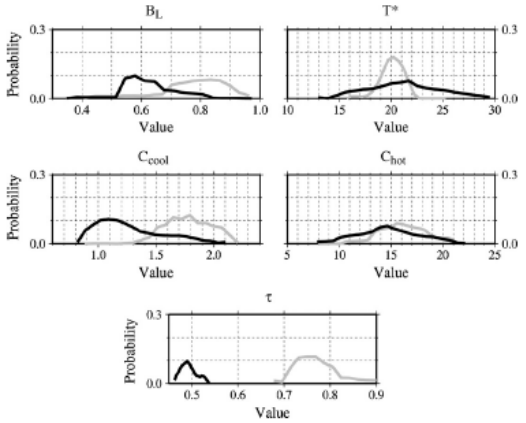


Fig. 6 – Posterior distributions of parameters for Aomori and Miyazaki. The gray line indicates Aomori, and the black one, Miyazaki. B_L: sensitivity of developmental rate to day length; T^*: base air temperature for calculating cooling degree days, C_{cool}: curvature factor of spikelet sterility caused by low temperature; C_{hot}: curvature factor of spikelet sterility caused by high temperature; and τ: technical coefficient.

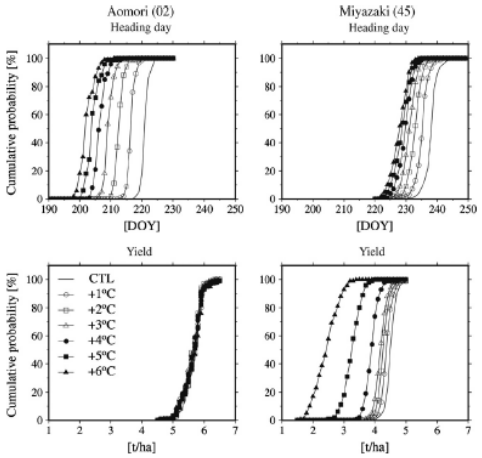


Fig. 7 – Cumulative probability density functions (CPDFs) of heading day and yield under six-level warming temperature conditions. They were estimated by the large-scale model with 5000 consistent sets of parameters obtained from the posterior distributions.

Is prediction error always the best criterion for model performance?

- Only if the model is being used for prediction
- We have discussed other goals of modeling (exploratory, explanatory, sensitivity analysis, etc.)
- In those cases, best model may not give best predictions
- There is an apparent tradeoff between model prediction and interpretation (i.e. best predictors are often multi-model averages) is this a surprise?