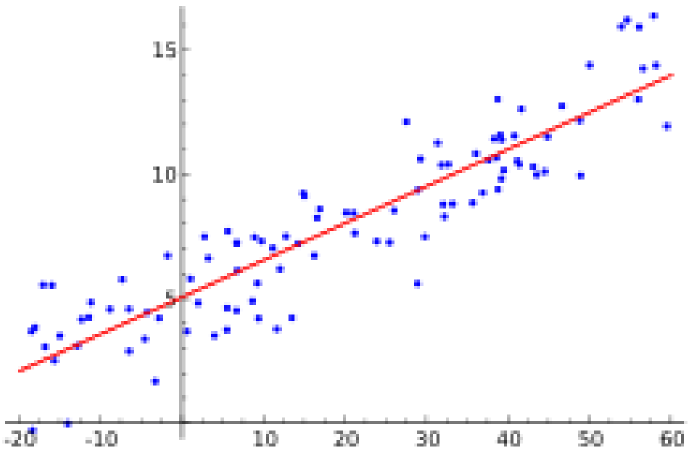
Logistic regression

Dr. Alessandro Filazzola

Generalized Linear Models (GLM)

- General linear models = typical linear regression

- Generalized Linear Models = distribution + link function
  - Poisson regression
  - Negative binomial regression
  - Gamma regression
  - Zero-inflated regression
  - Logistic regression (sometimes referred to as binomial)



Why use a GLM?

1) The response variable is not normally distributed, or the range is restricted

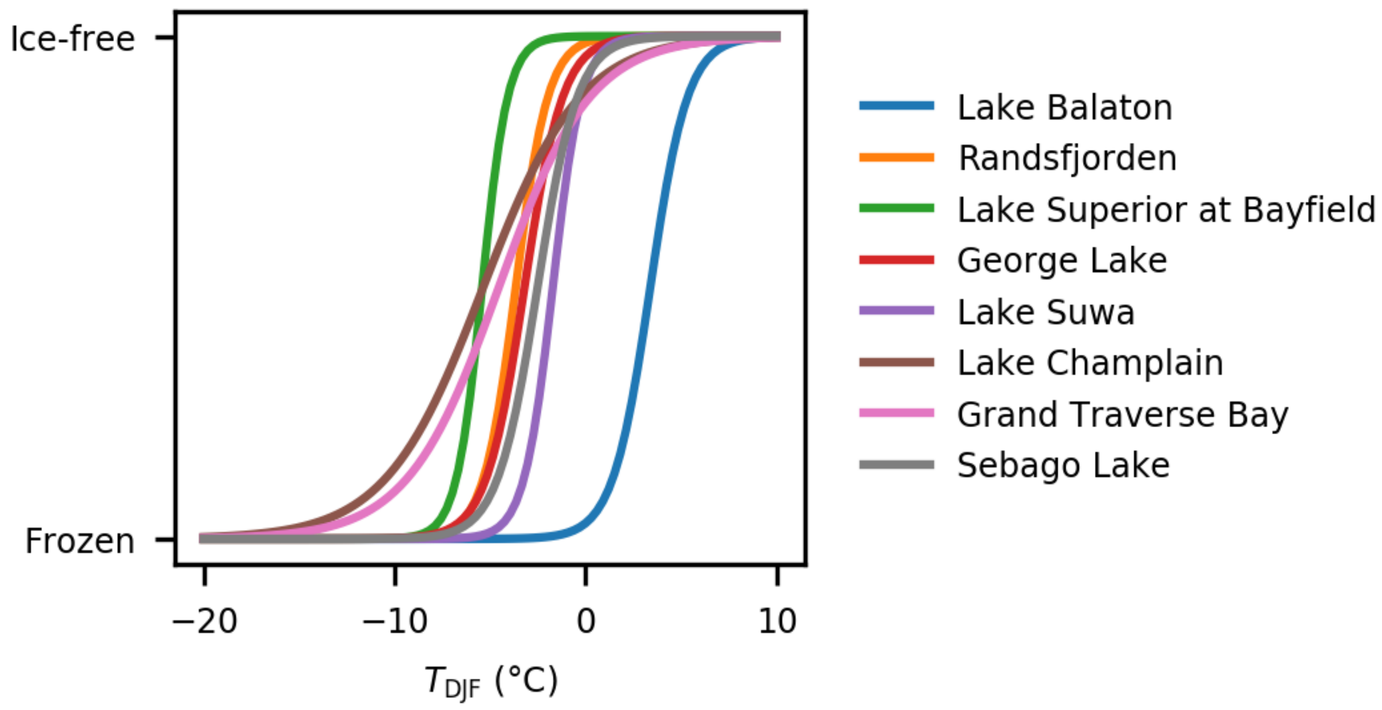
2) The variance of the response variable depends on the mean

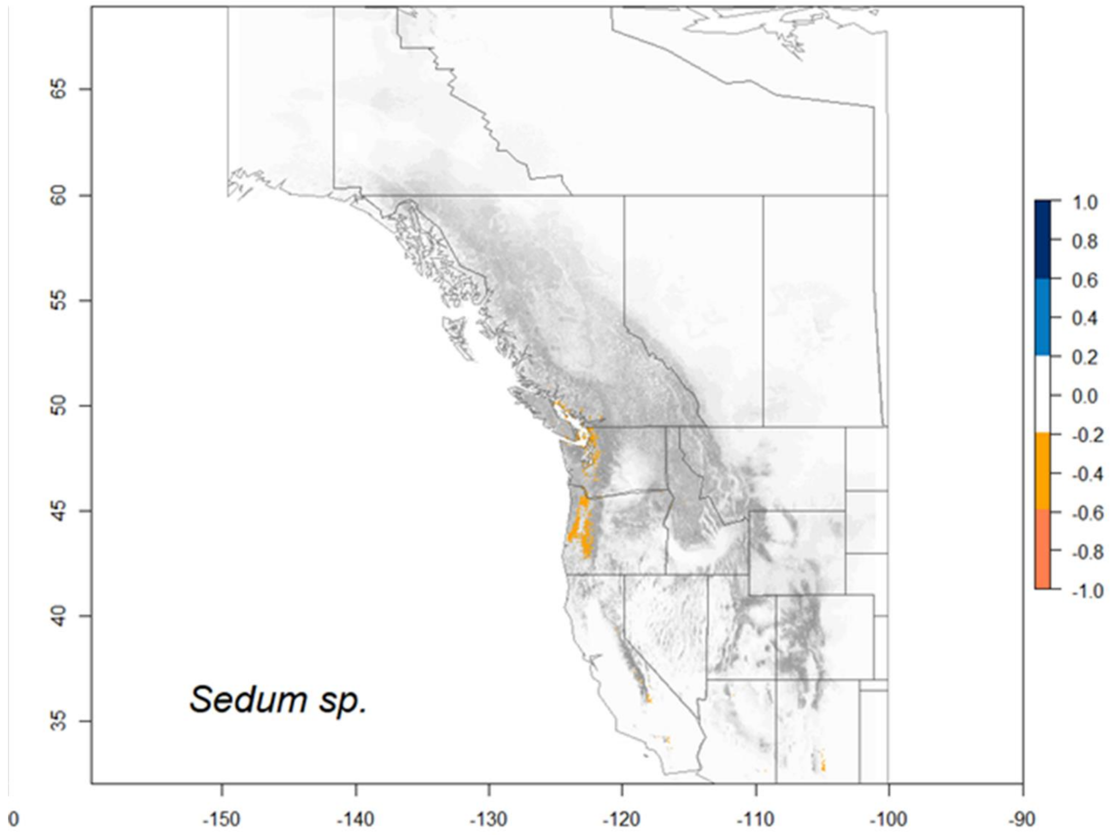
3) GLMs can fit to particular distributions without transformations. The benefits of this include i) the homogeneity of variance does NOT need to be satisfied and ii) the response variable does not need to be changed.

$$E(\log(Y)) \neq \log(E(Y))$$

Logistic regression

Used frequently to **test** and **predict**

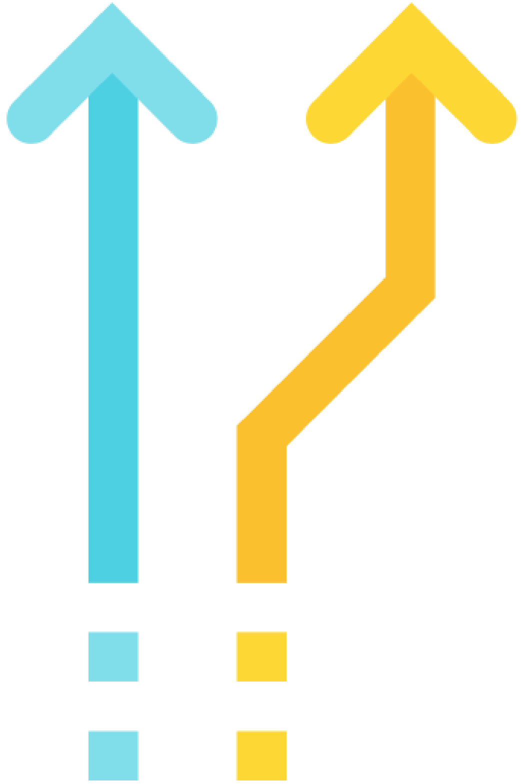




Parallels with linear regression

Both logistic and linear regression can:

- Test for significance of a predictor
- Calculate an effect size
- Use continuous or categorical predictors
- Can predict values for new values



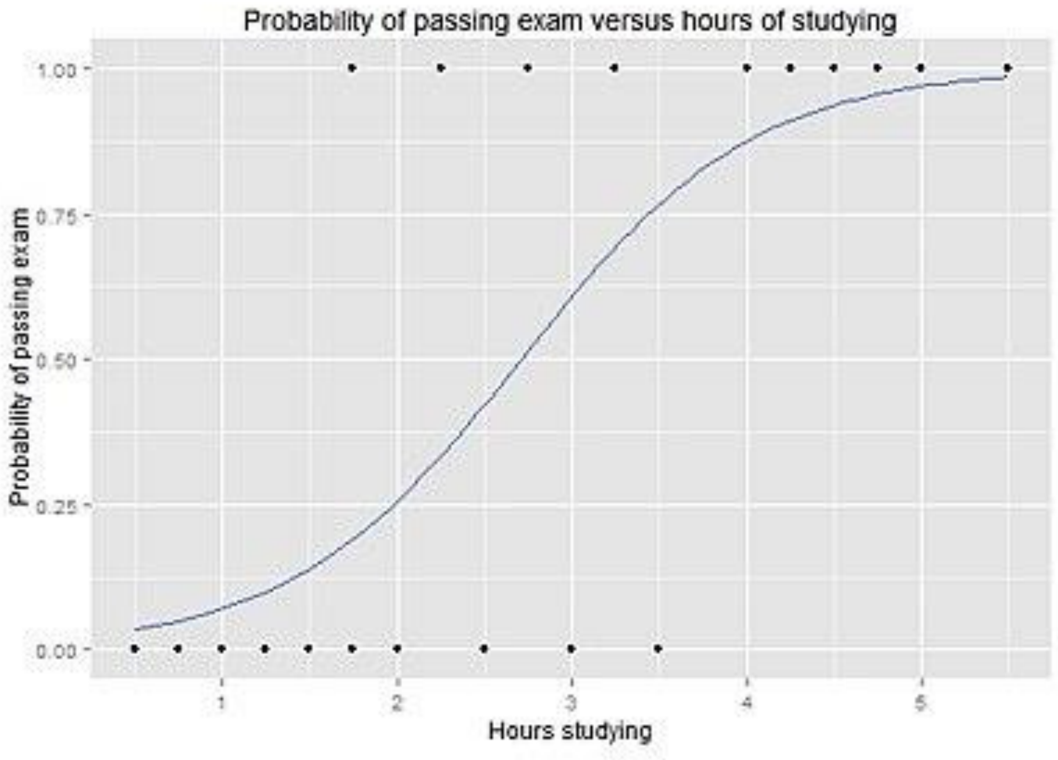
Differences with linear regression

Logistic regression differs from linear regression in that:

- Binary outcome
- There is no “true” R2 or residuals
- Uses Maximum Likelihood instead of Sum of Squares

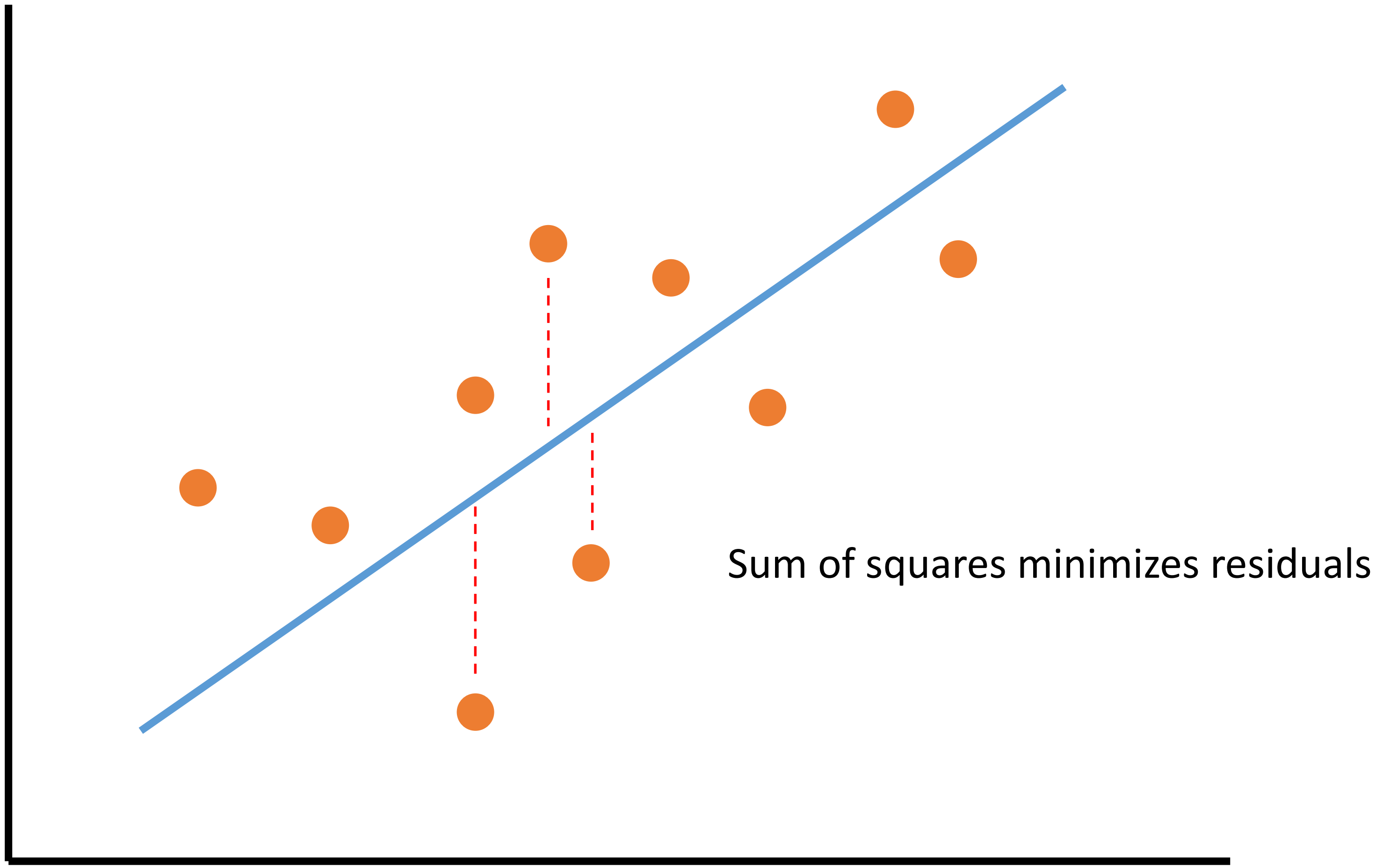
Understanding logistic regression

1) Fitting a line
2) The outputs
3) Calculating fit
4) Prediction

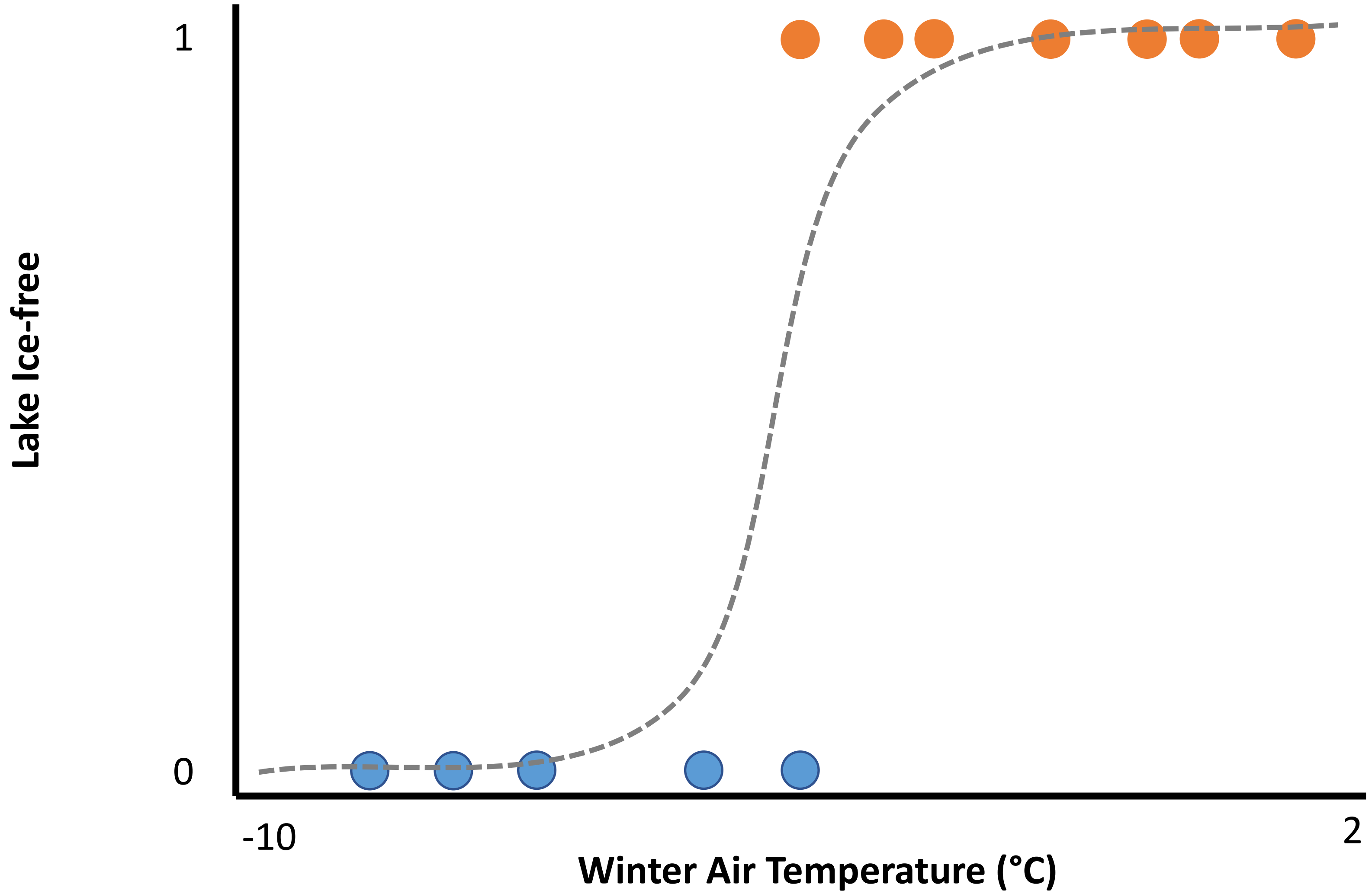


Wikimedia.com

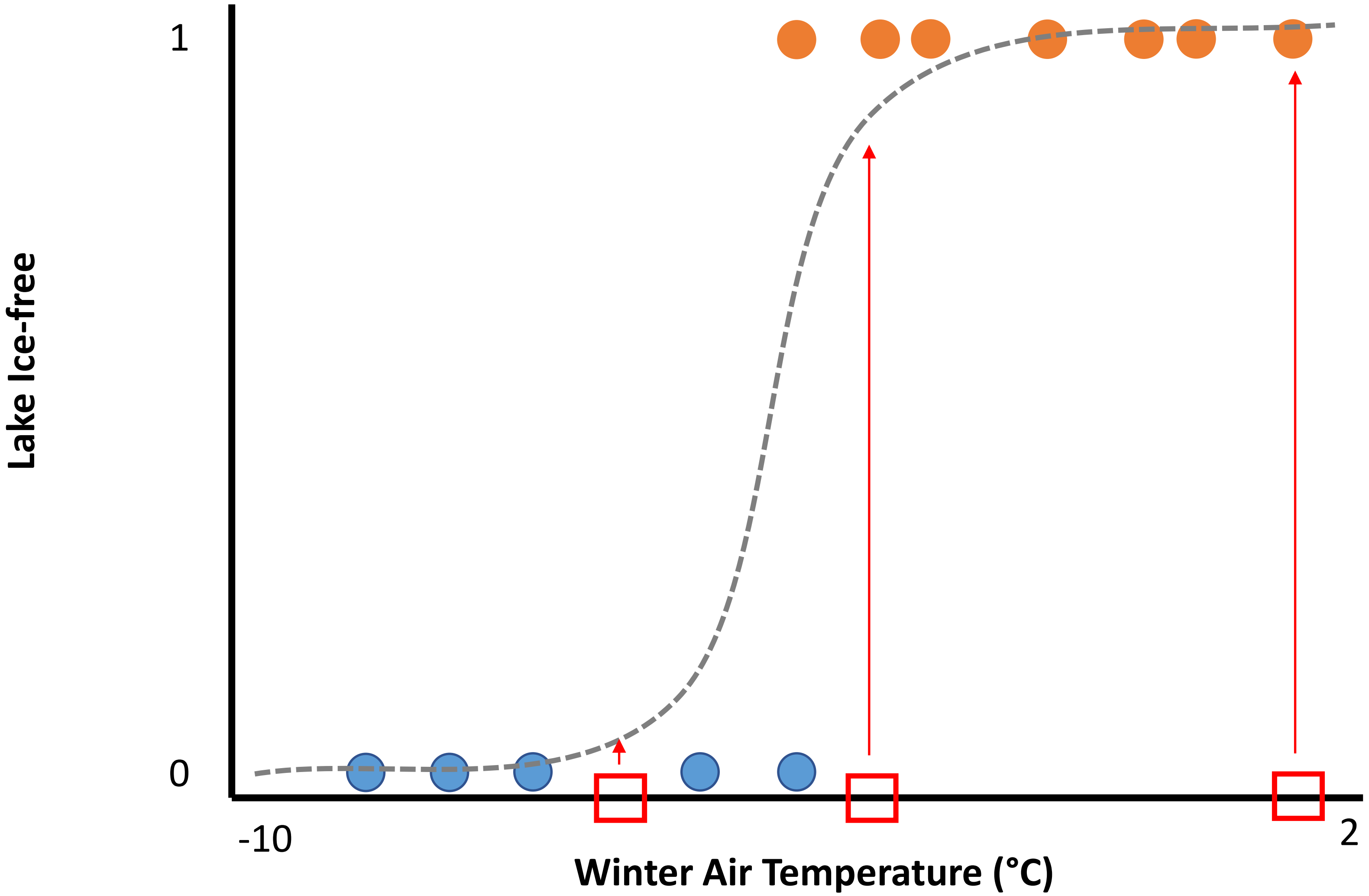
1) Fitting a line: linear regression



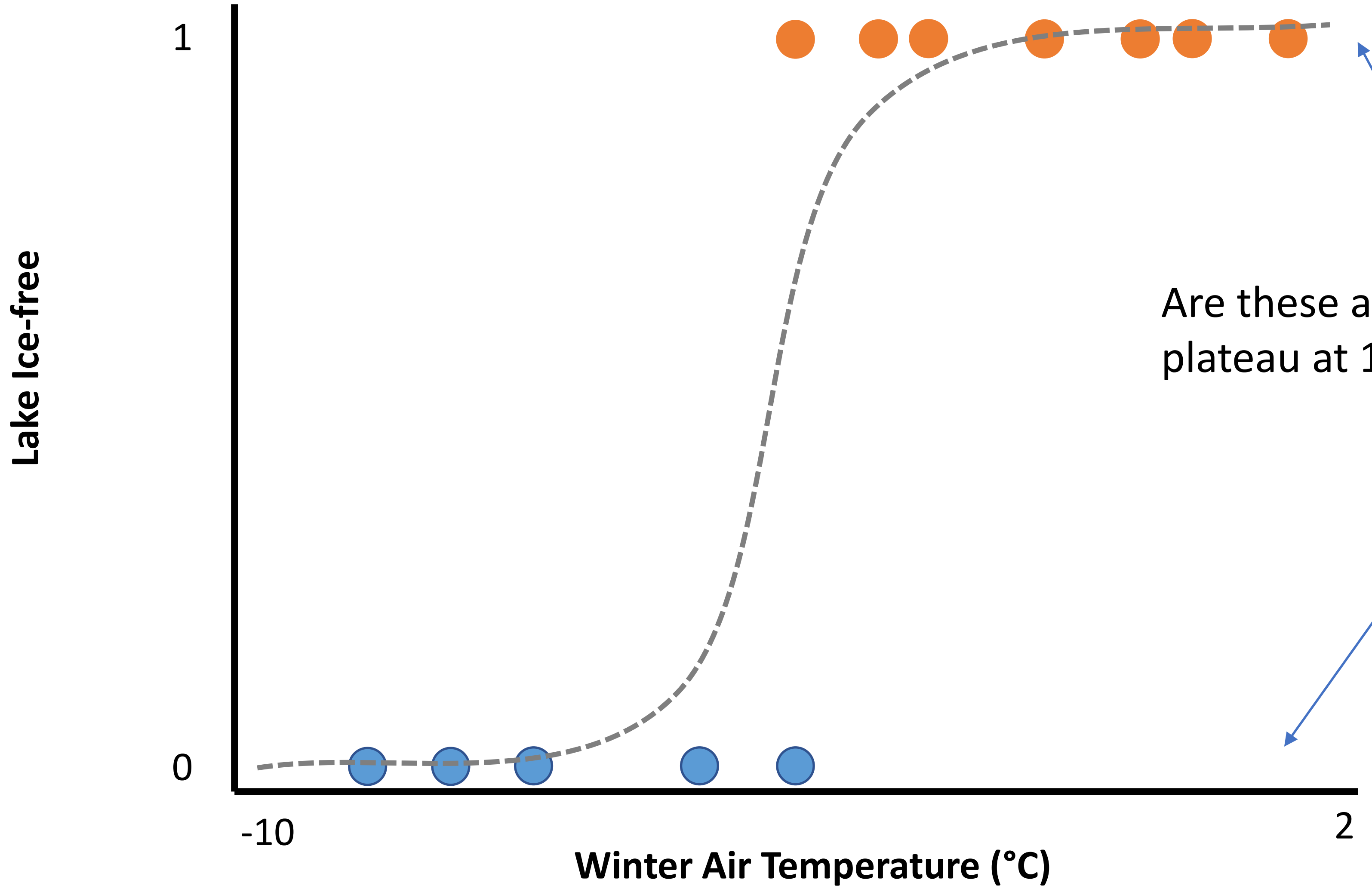
1) Fitting a line: logistic regression



1) Fitting a line: logistic regression



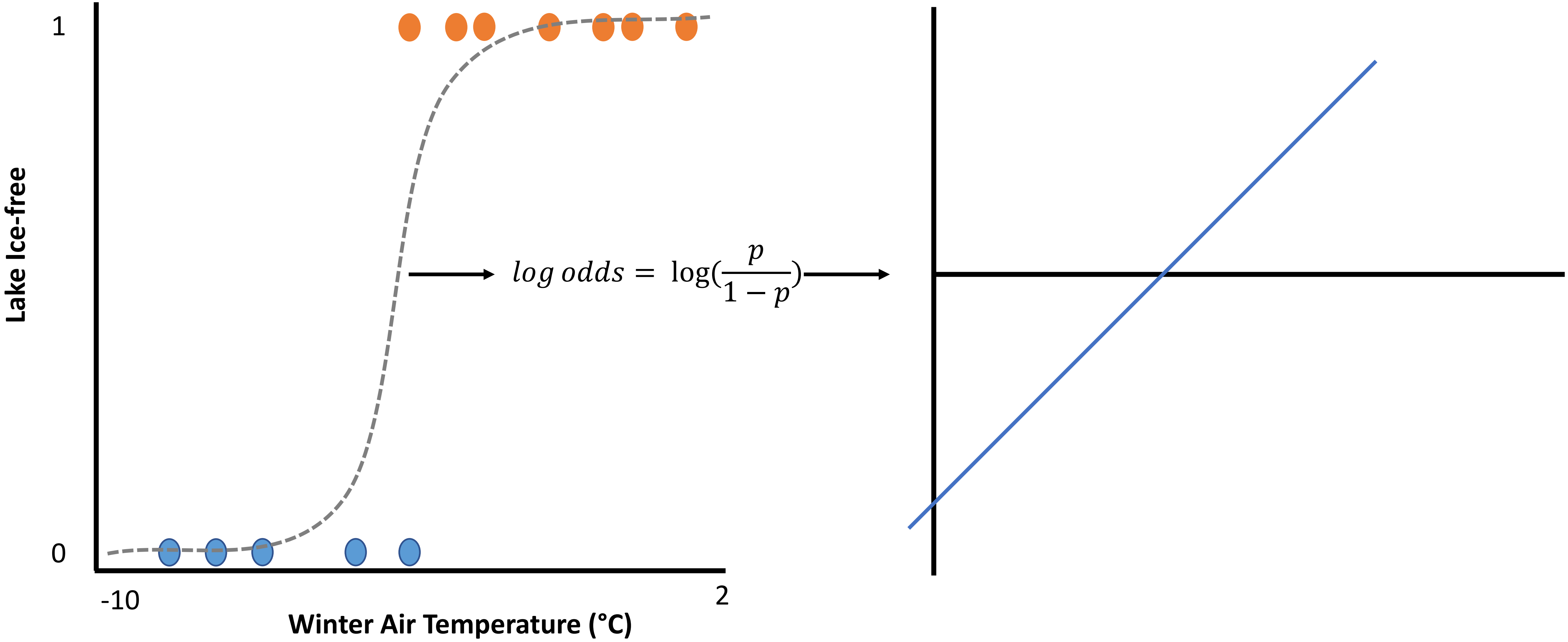
Question



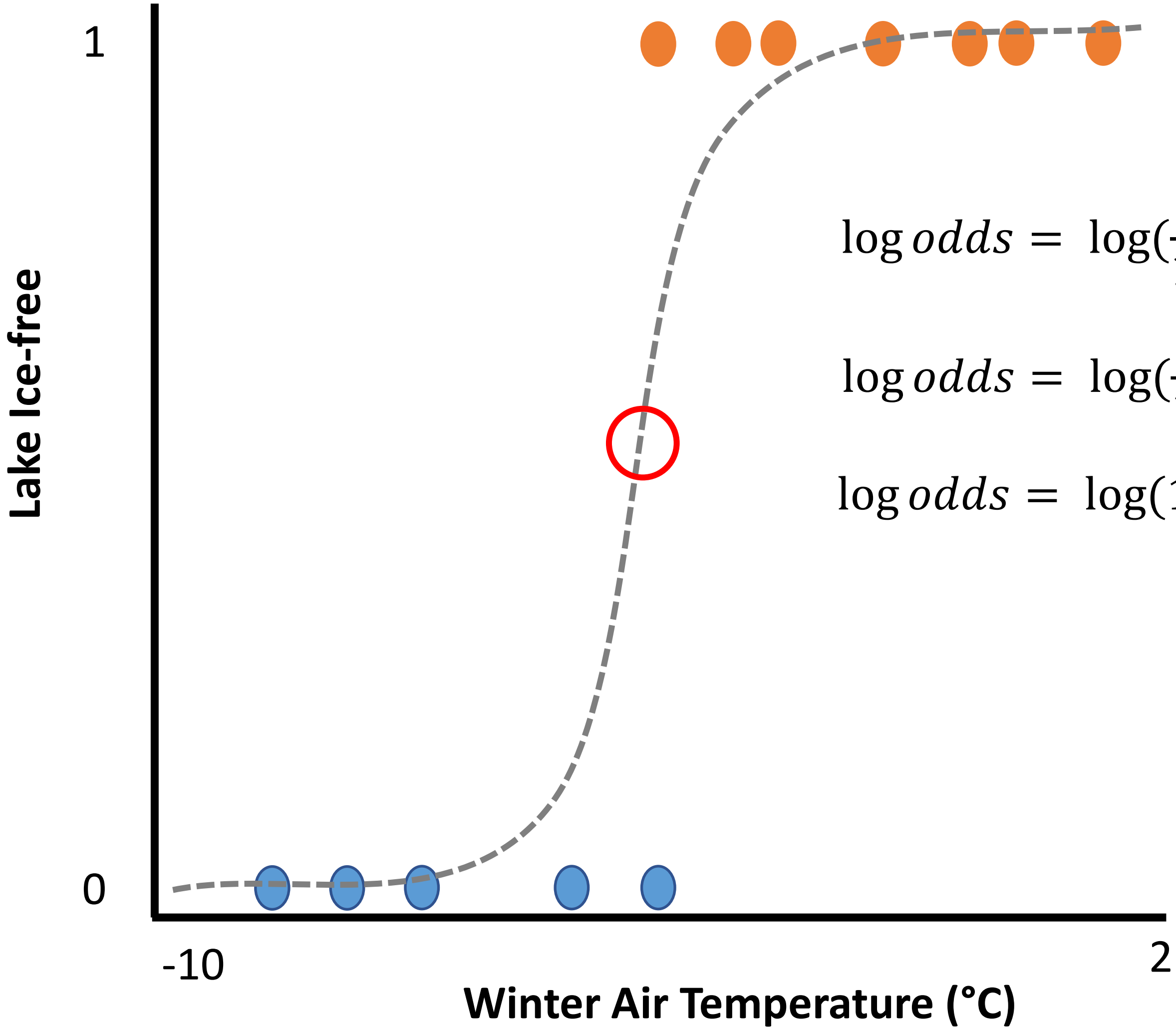
Are these asymptotes or to they plateau at 1 and zero respectively?

$$log\ odds = \log\left(\frac{p}{1-p}\right)$$

1) Putting the *linear* in generalized linear model



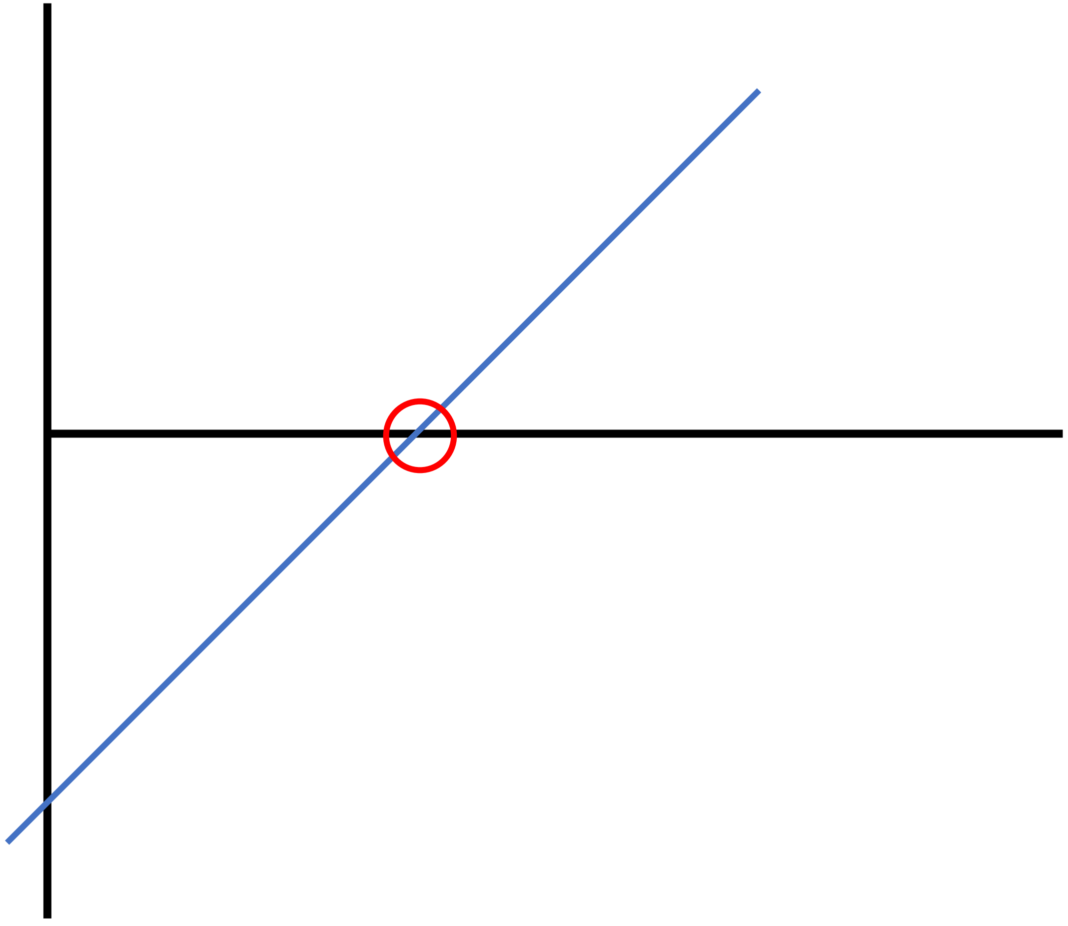
1) Putting the *linear* in generalized linear model



$$\log odds = \log(\frac{p}{1-p})$$

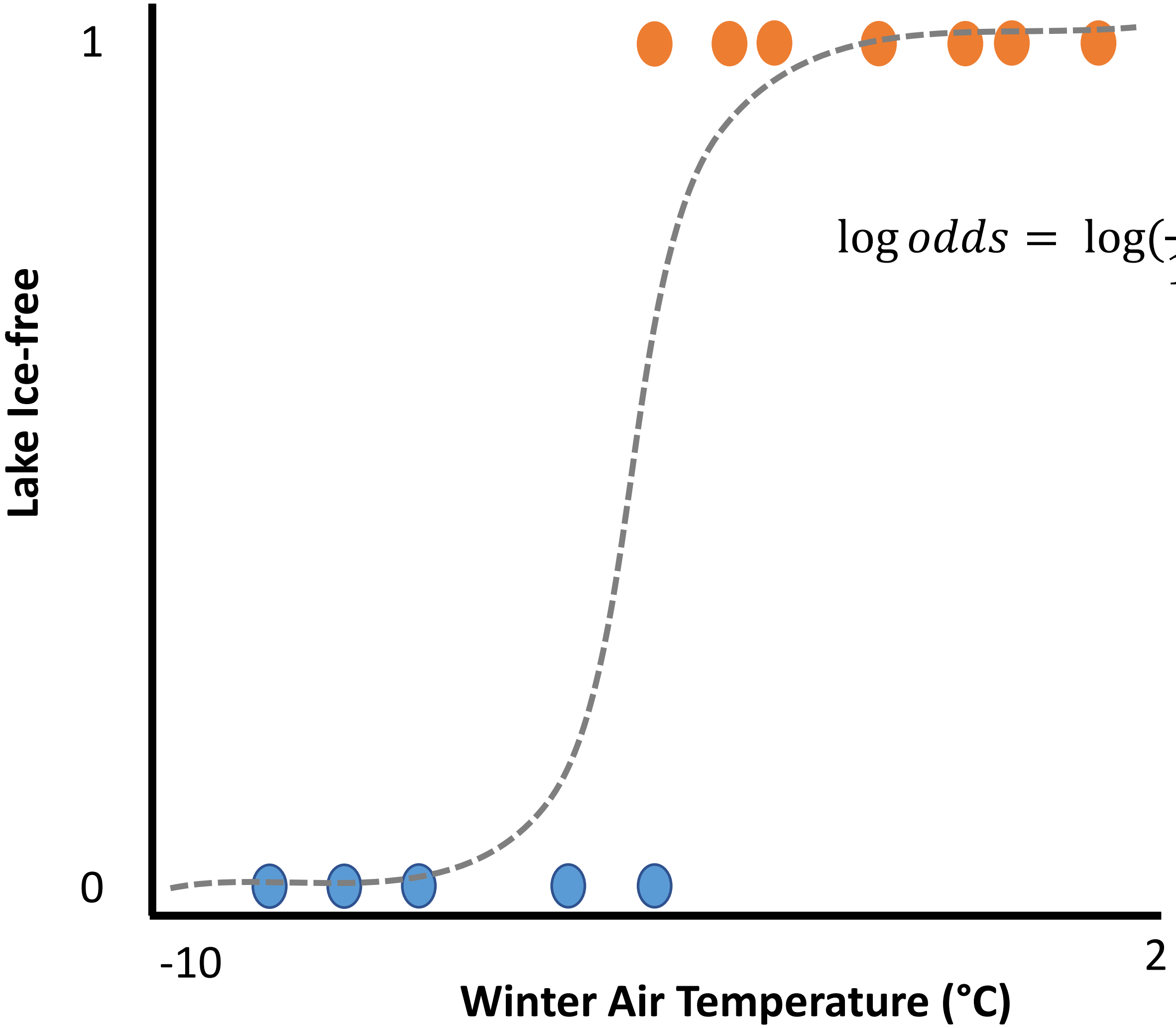
$$\log odds = \log(\frac{0.5}{1-0.5})$$

$$\log odds = \log(1)$$

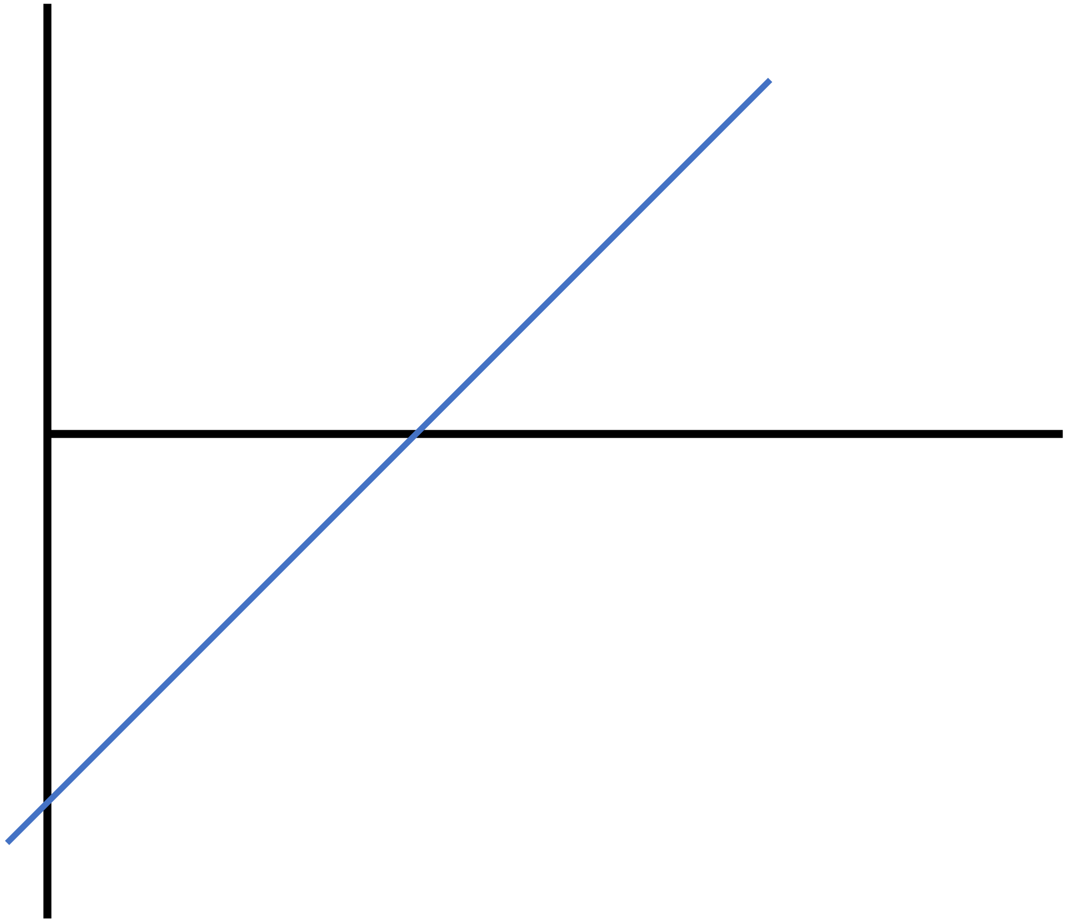


This is the slope co-efficient from the R Output

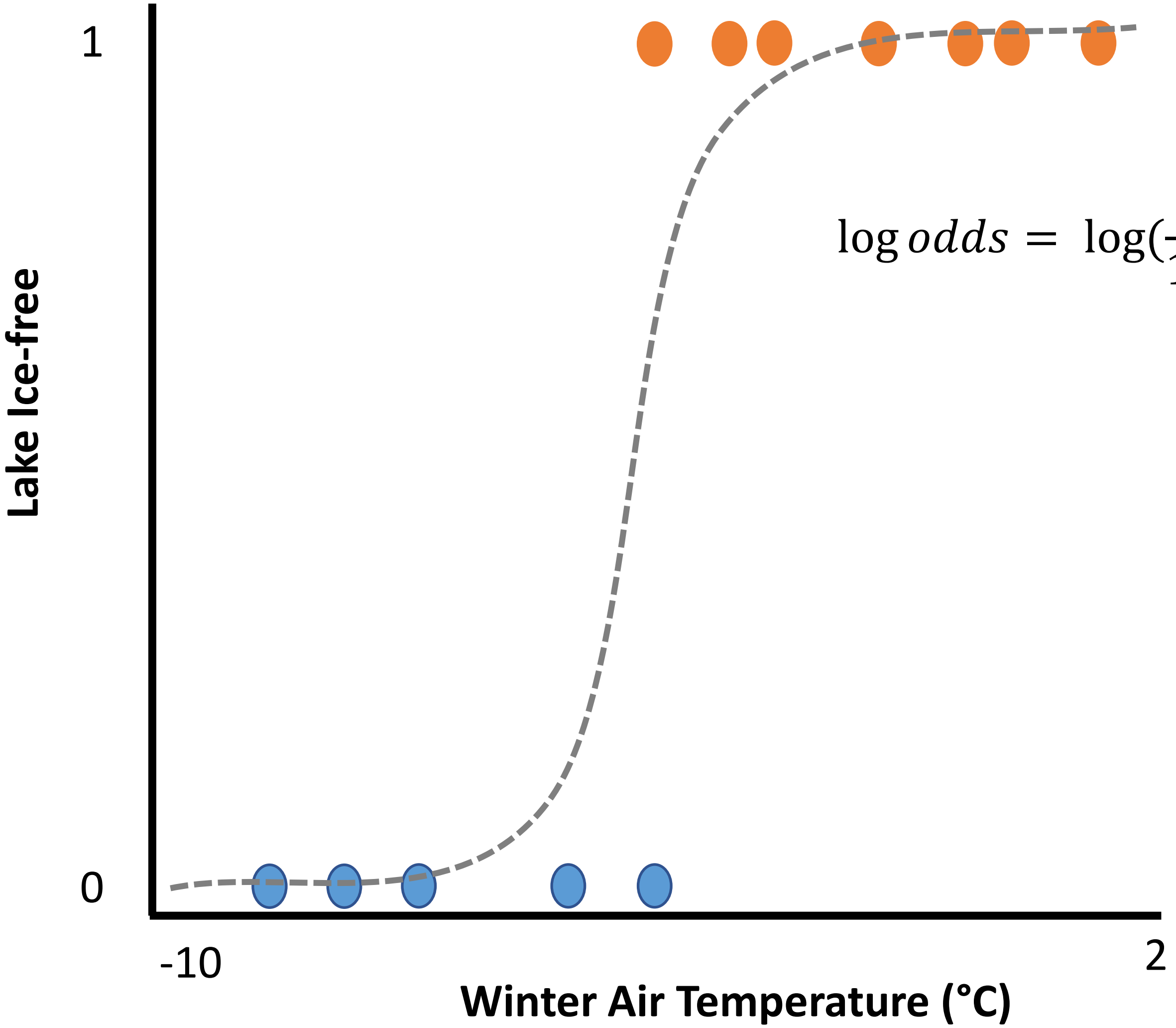
1) Putting the *linear* in generalized linear model



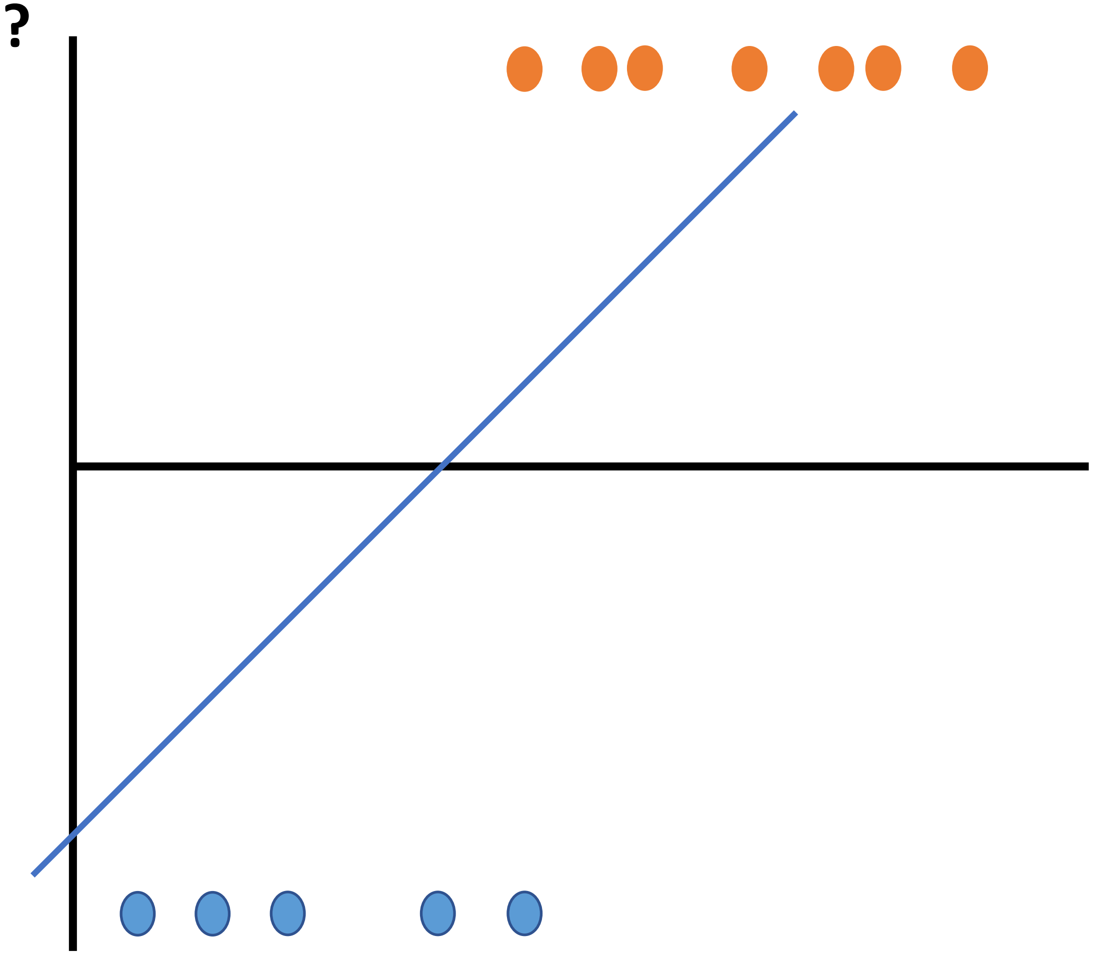
$$\log odds = \log\left(\frac{p}{1-p}\right)$$



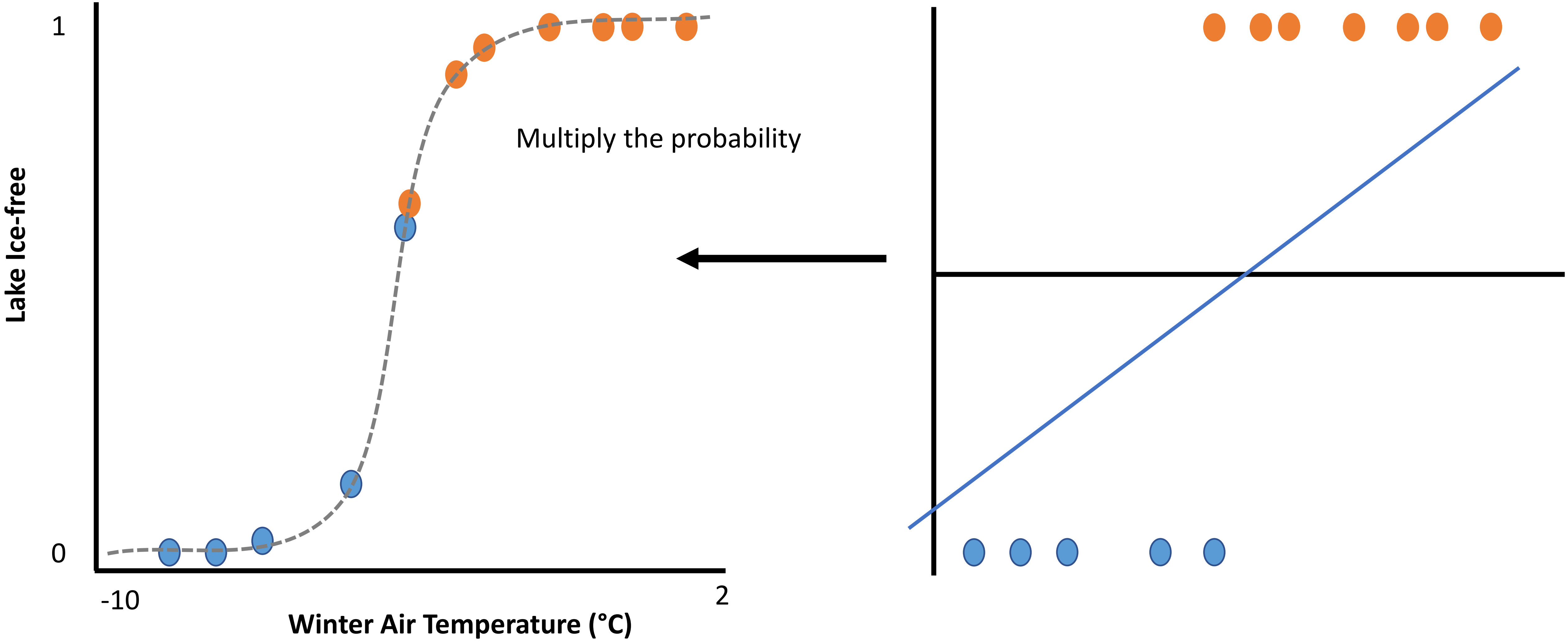
1) Putting the *linear* in generalized linear model



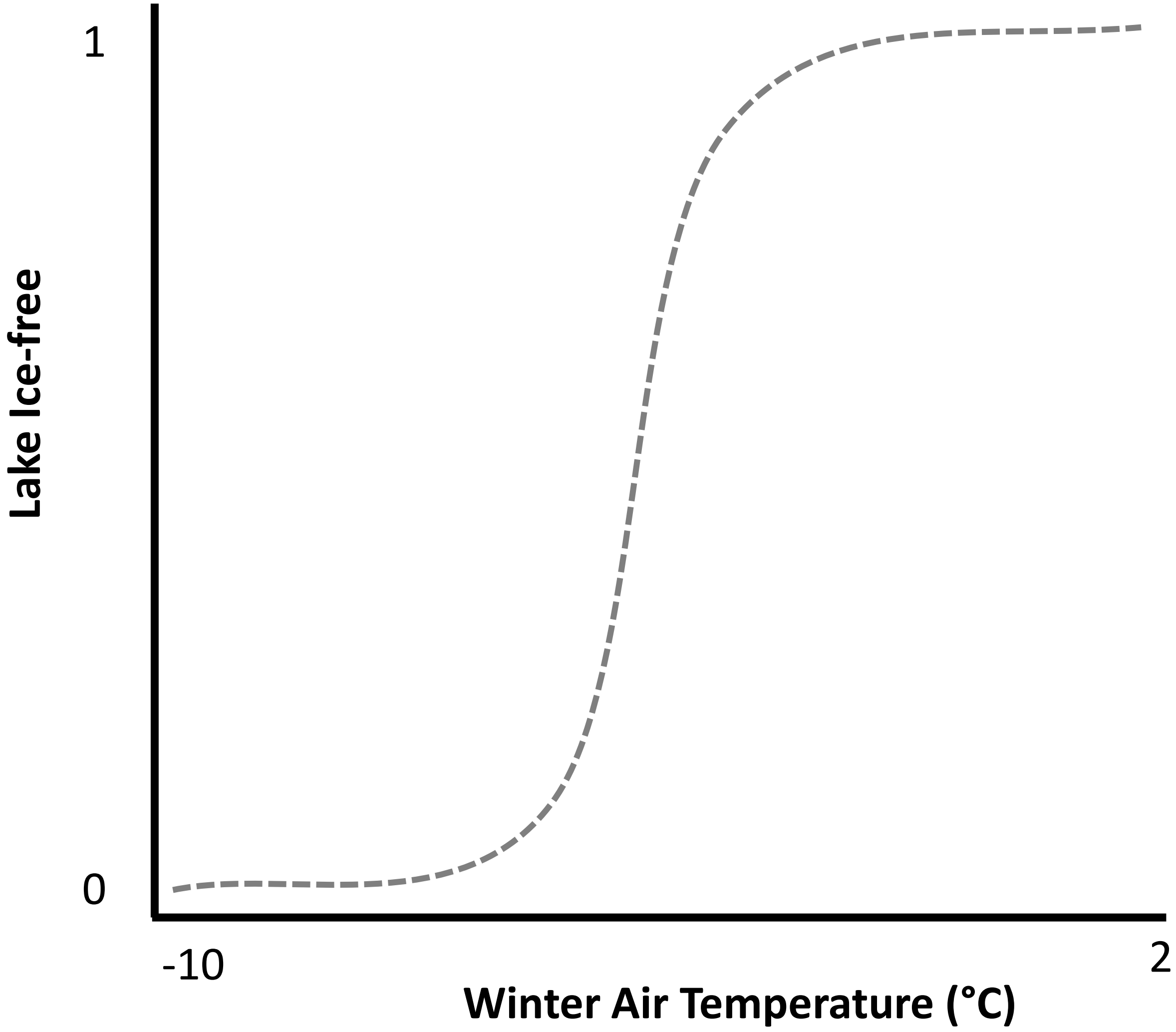
$$\log odds = \log\left(\frac{p}{1-p}\right)$$

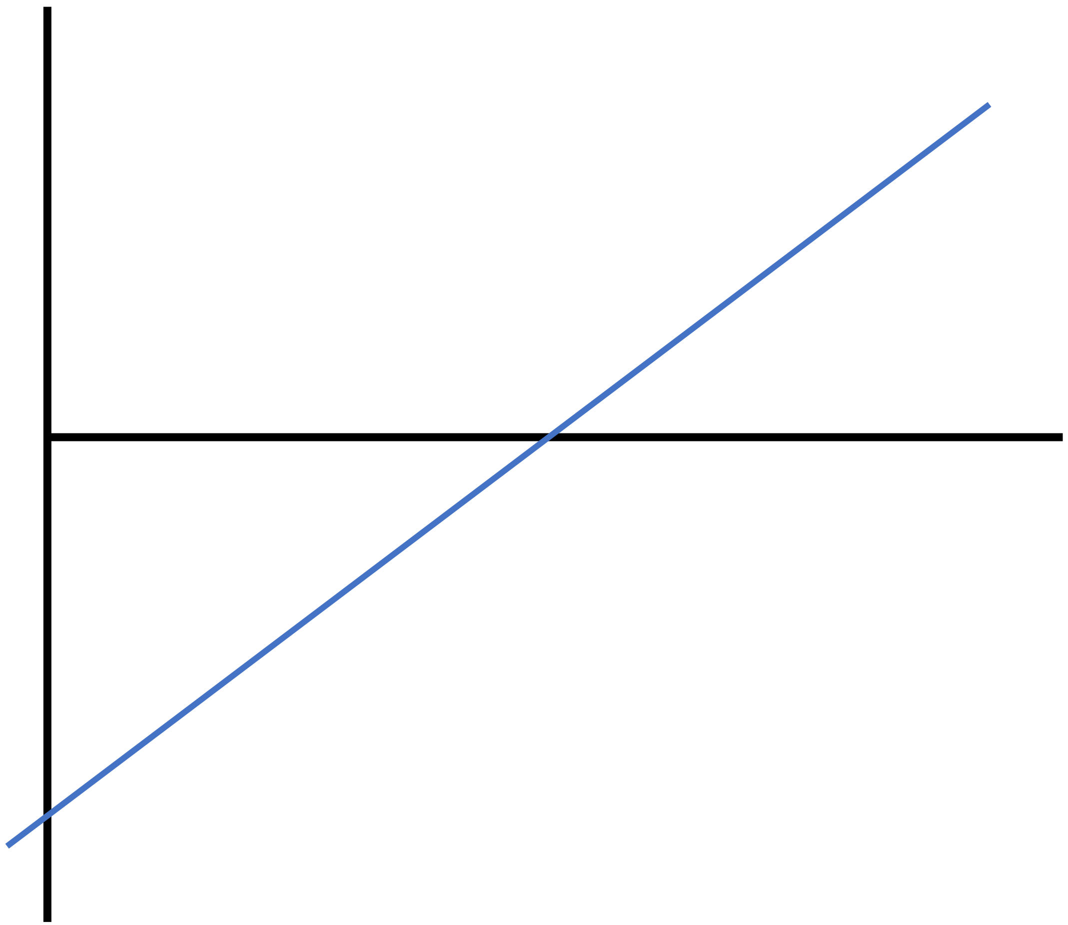


1) Fitting the model

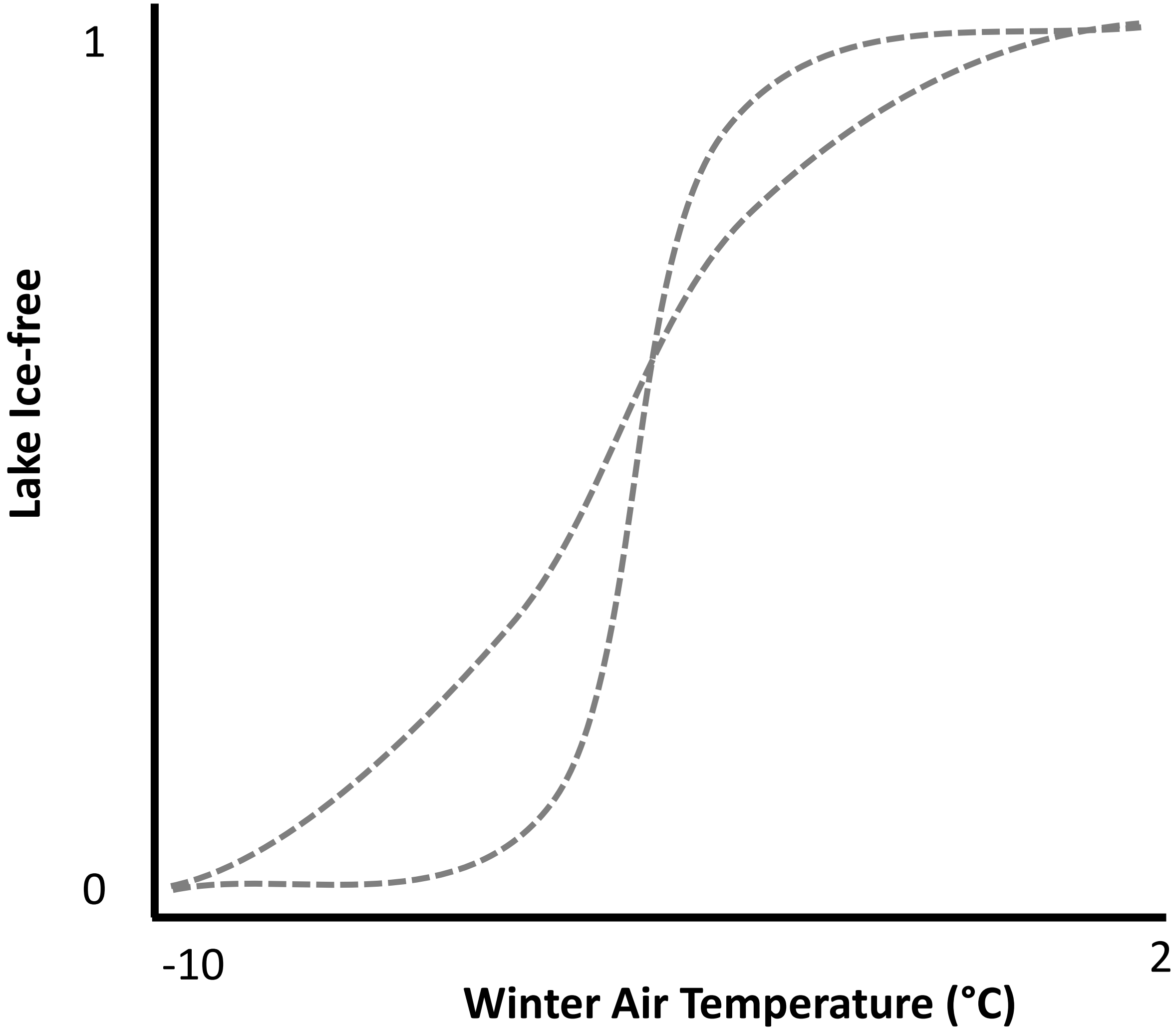


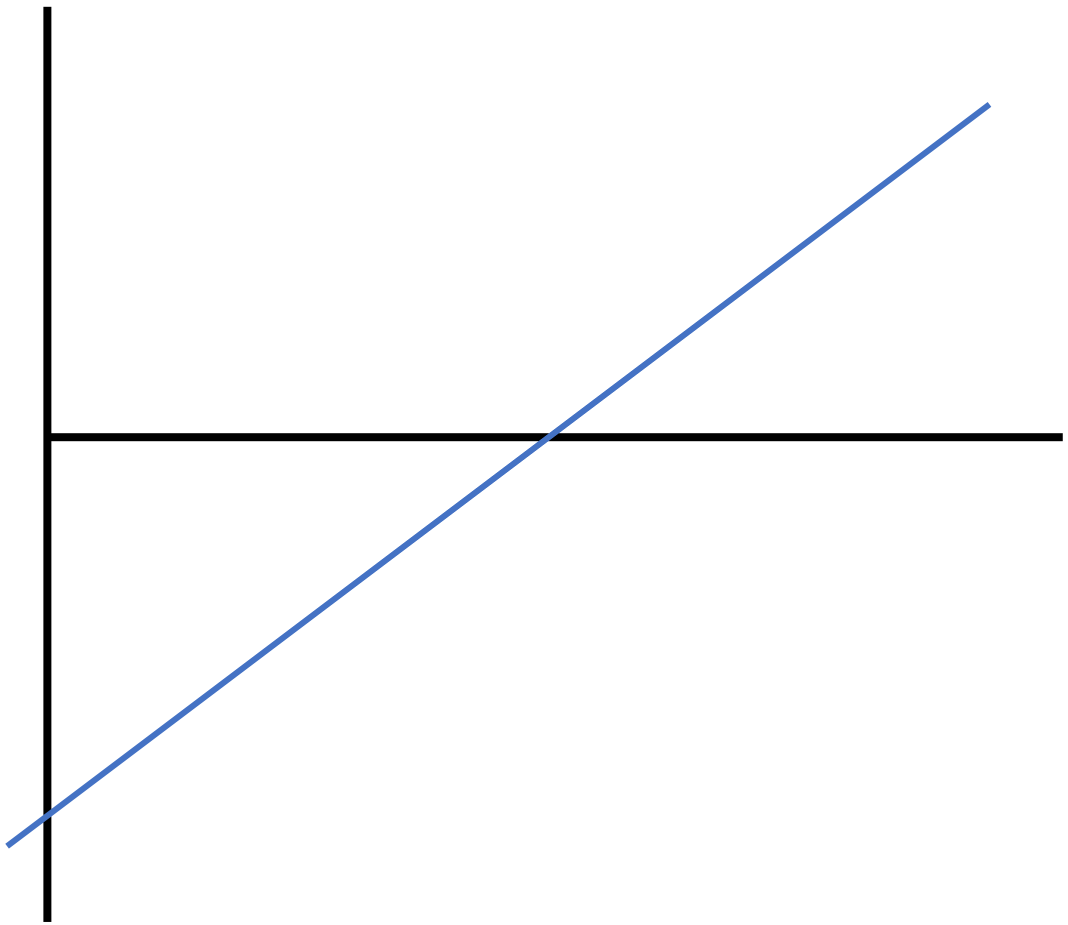
1) Fitting the model – maximum likelihood



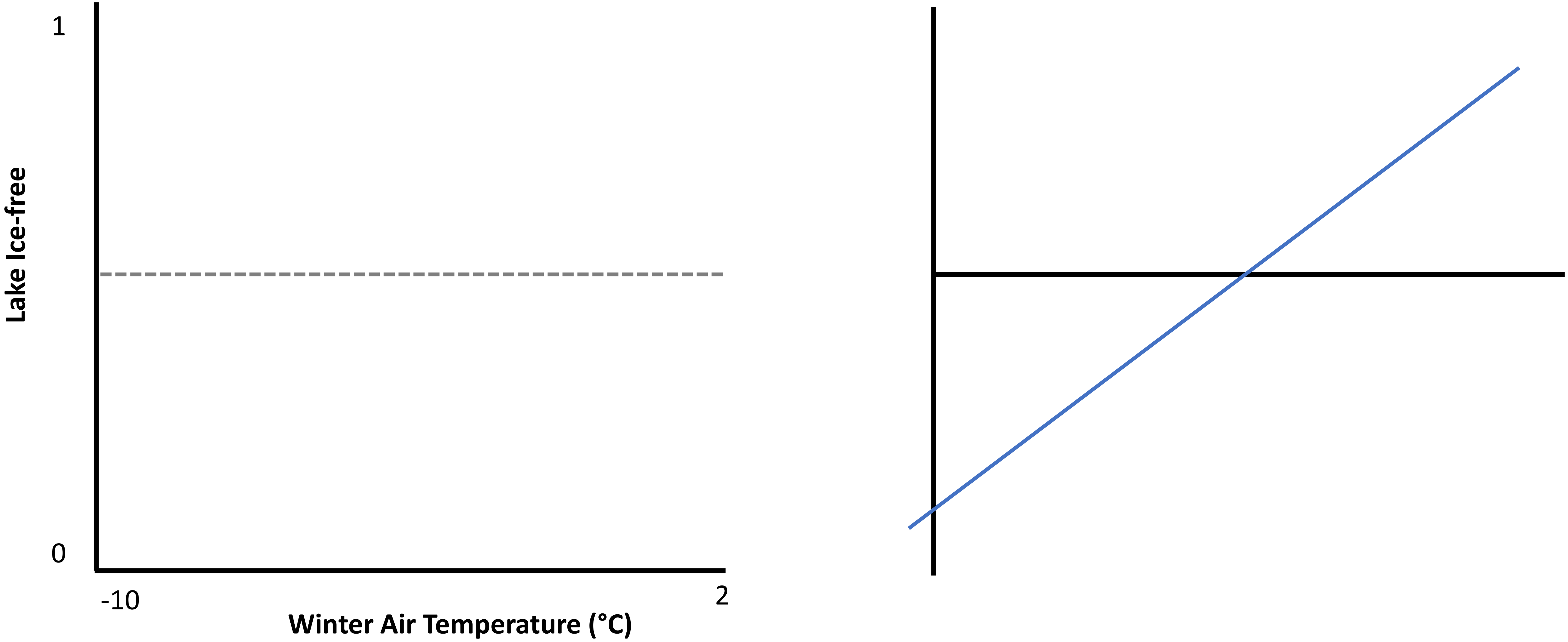


1) Fitting the model – maximum likelihood

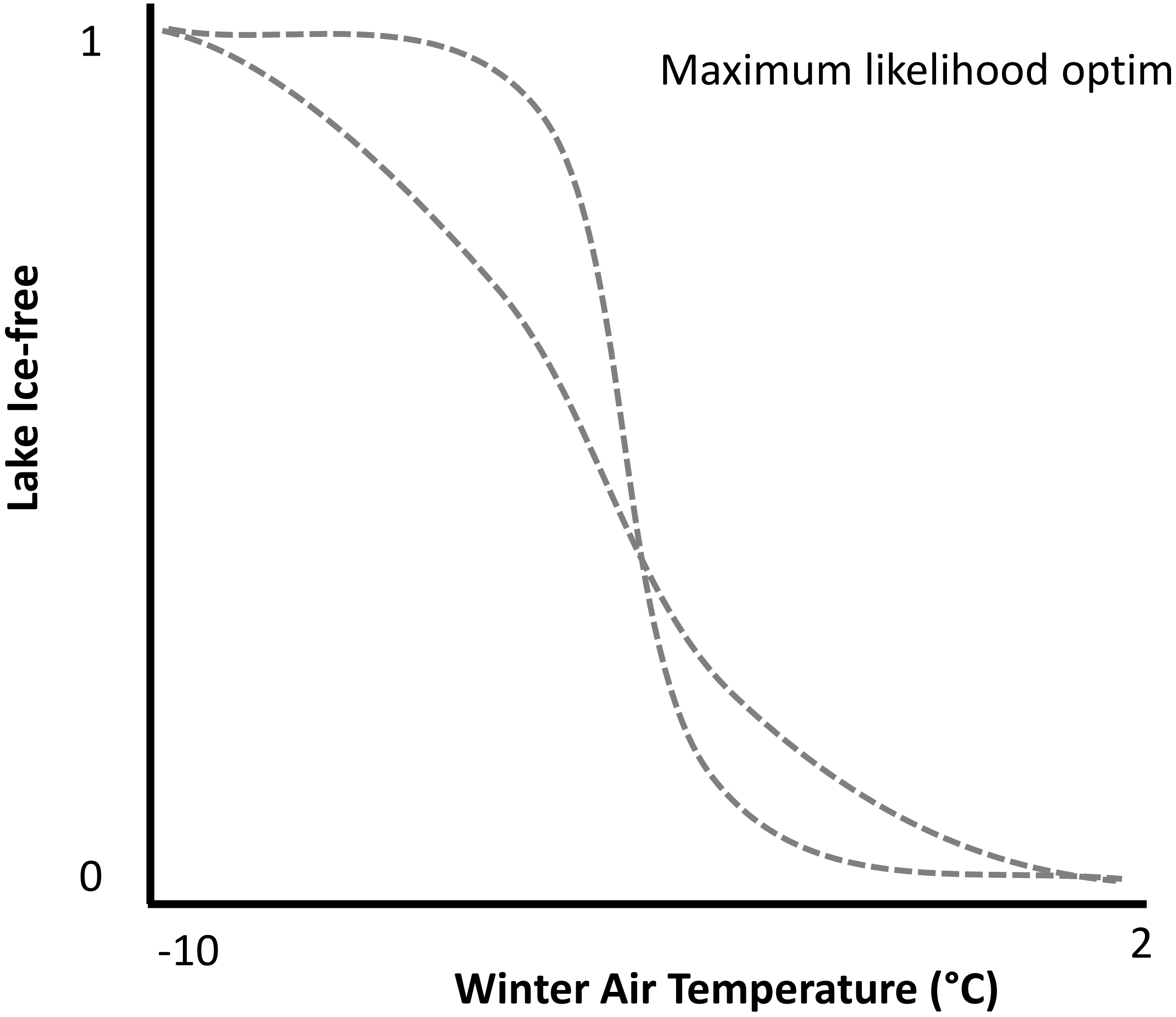




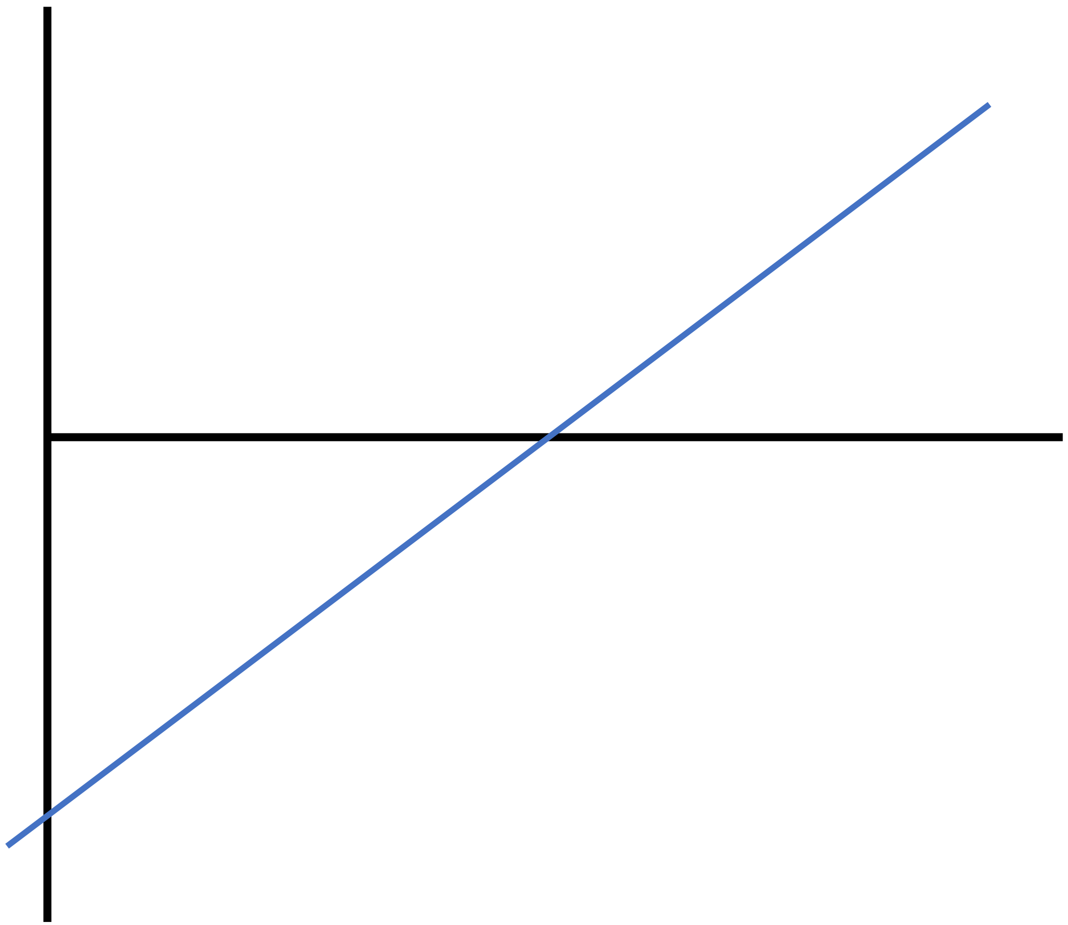
1) Fitting the model – maximum likelihood



1) Fitting the model – maximum likelihood



Maximum likelihood optimizes quickly



2) The outputs - Logistic regression in R

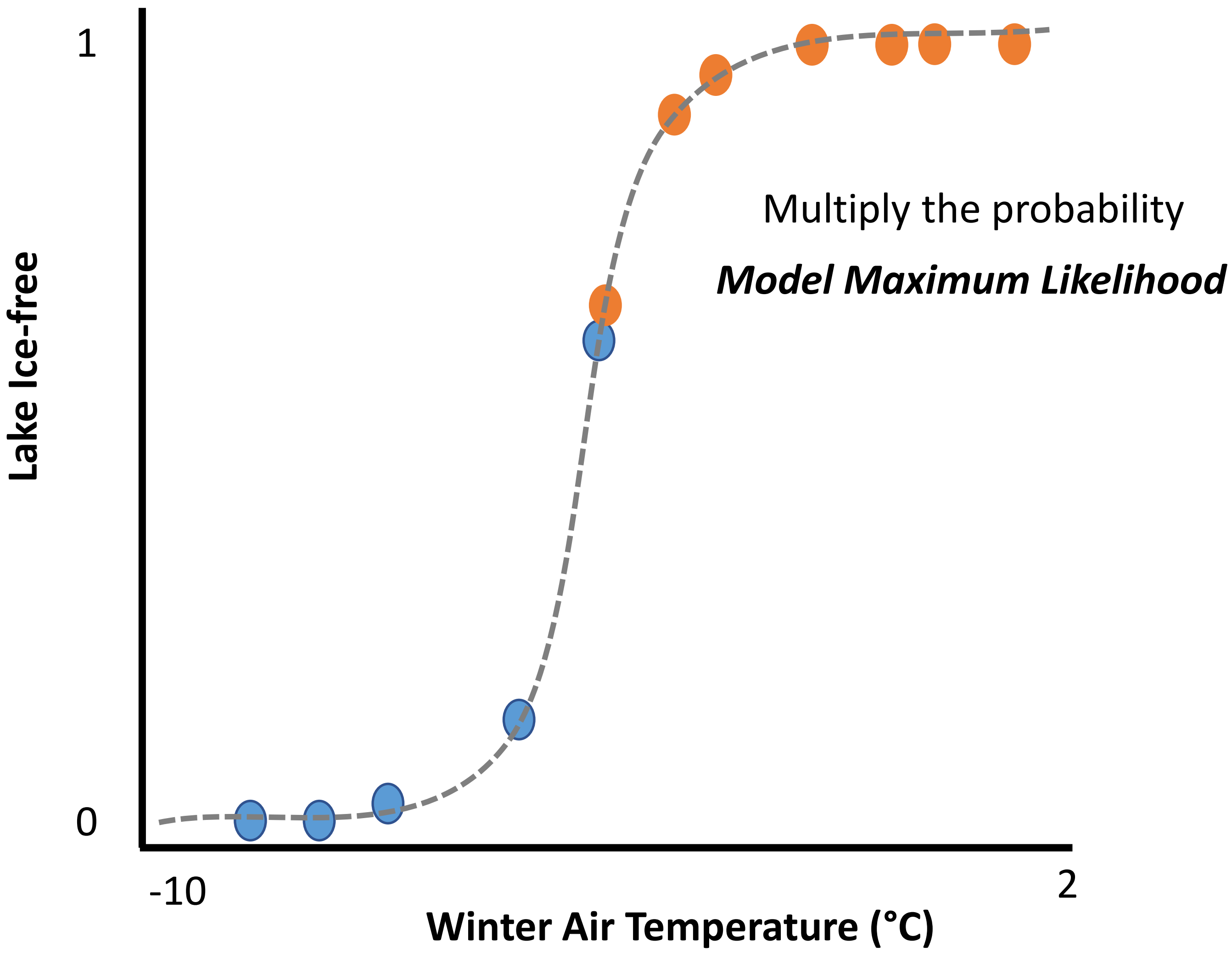
3) Calculating a measure of fit

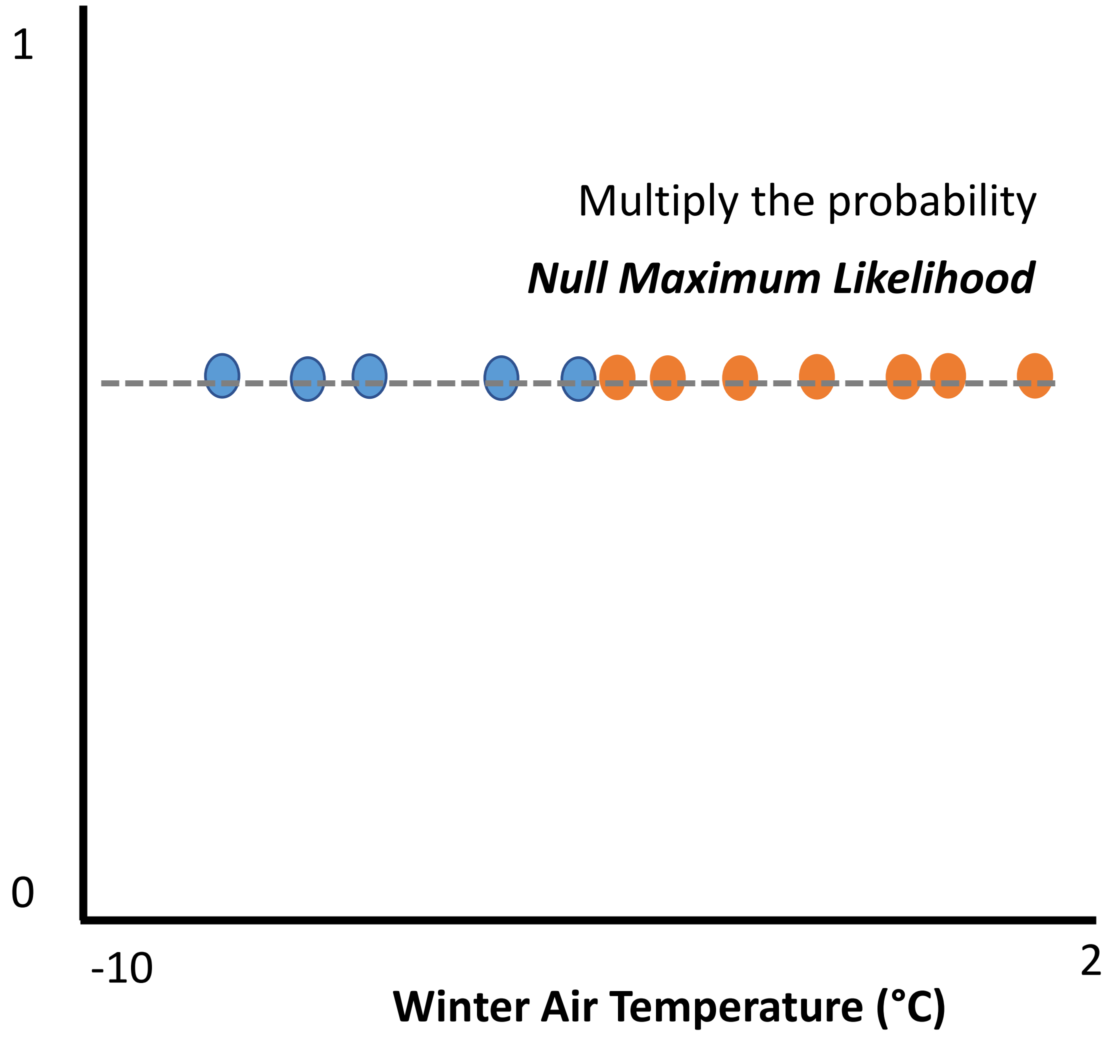
- Typical linear regression uses R2

- No R2 in logistic regression because no residuals

- Instead use Psuedo R2
  - McFadden’s among the simpler compares to null

3) Calculating a measure of fit





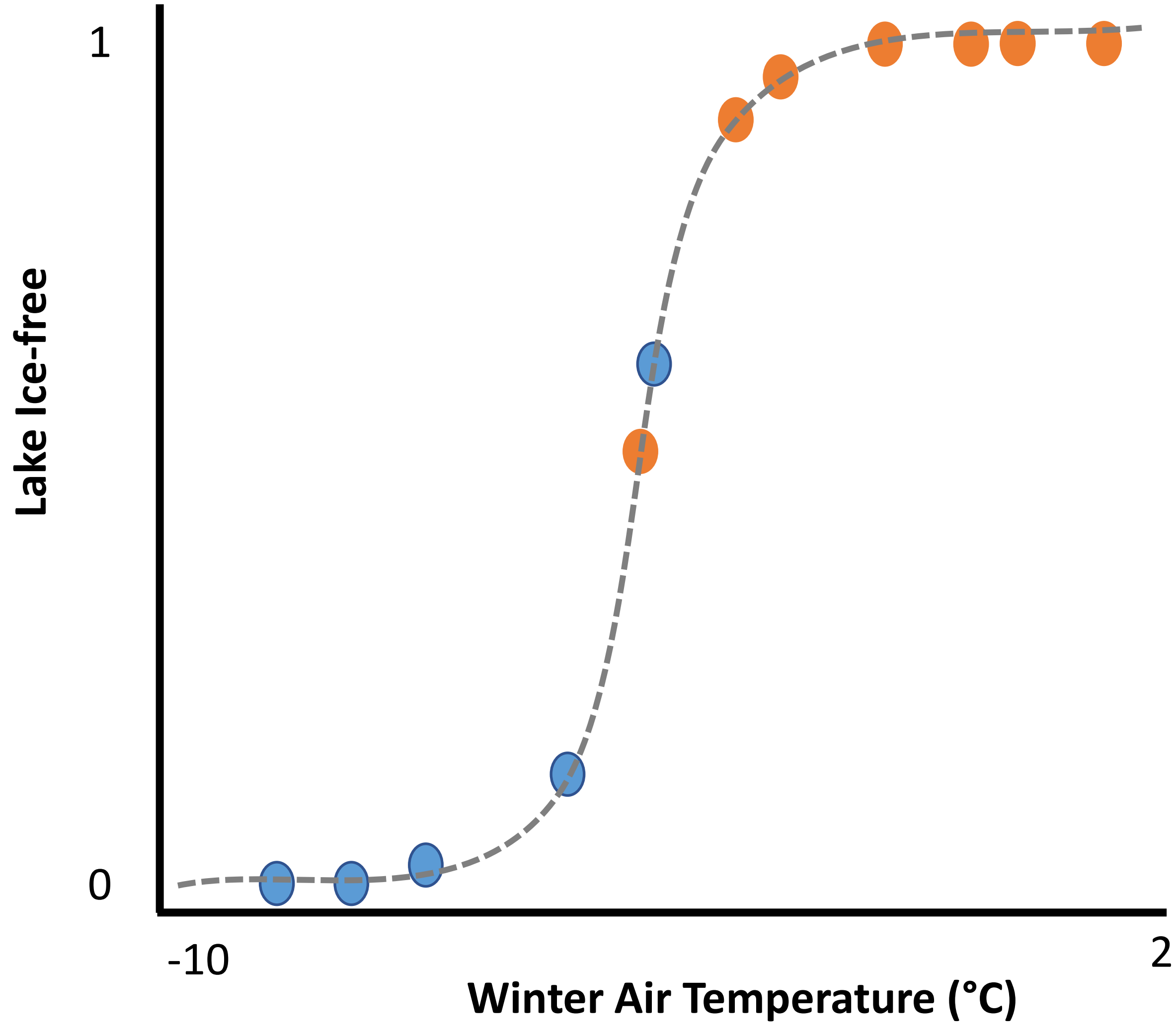
3) Calculating a measure of fit

$$\text{PsuedoR}^2 = 1 - \frac{\log(Model)}{\log(Null)}$$

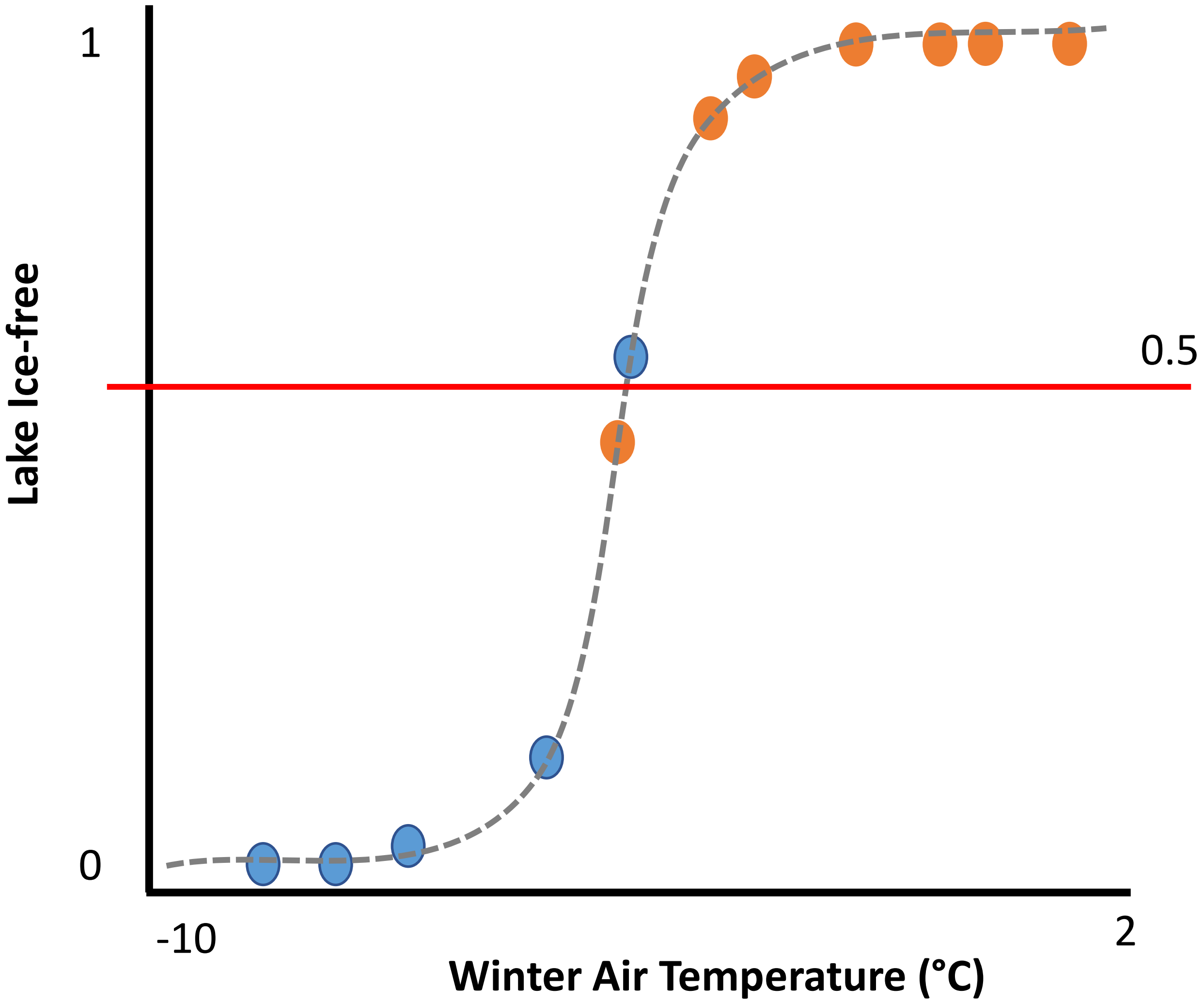
```
m3<- glm(icefree~temperature,  family="binomial", data=iceData)
m3null<- glm(icefree~1,  family="binomial", data=iceData)
1-logLik(m3)/logLik(m3null)
```

4) Prediction

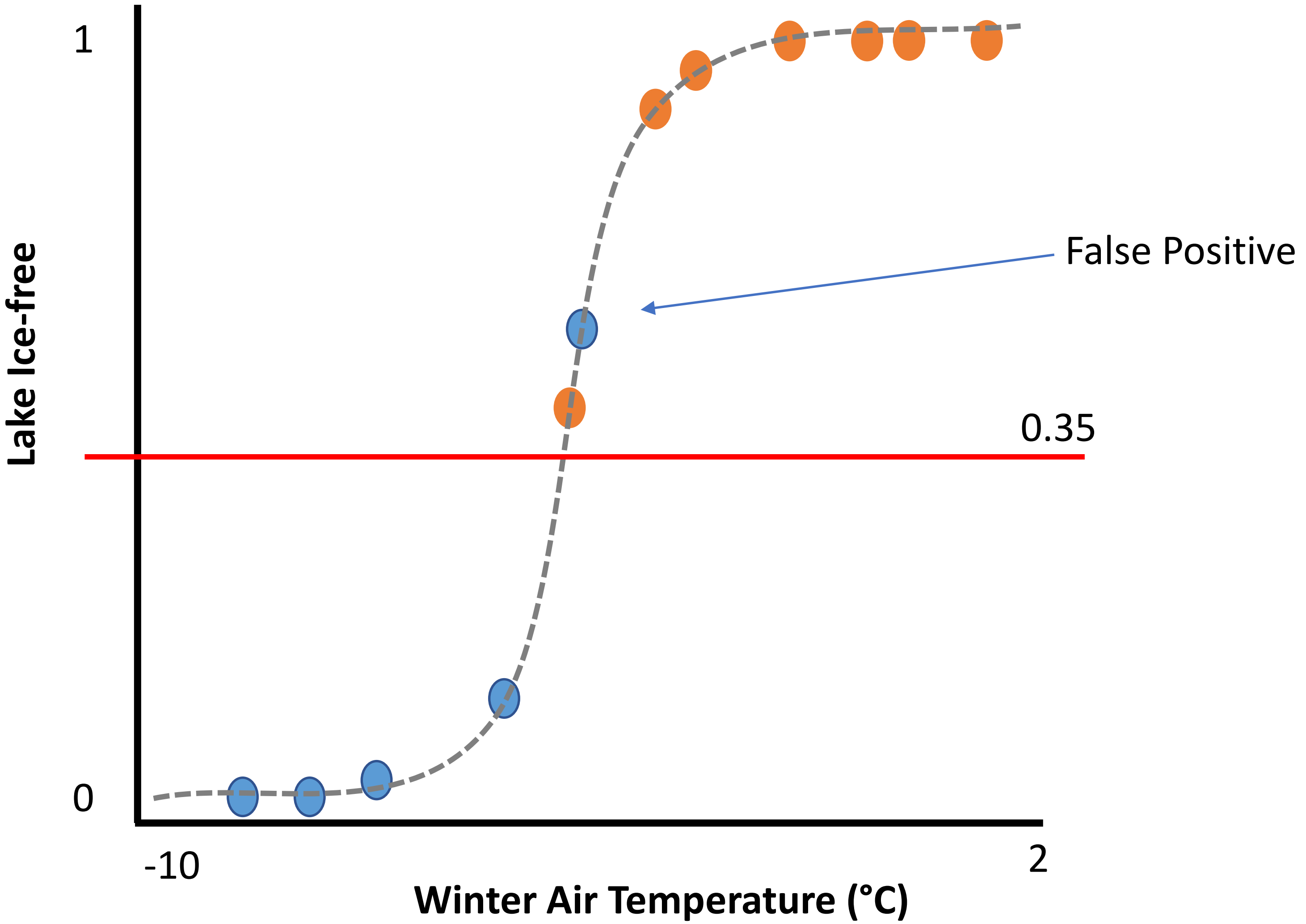
- Models probability
- ..but the outcome is binary
- Need to identify a threshold



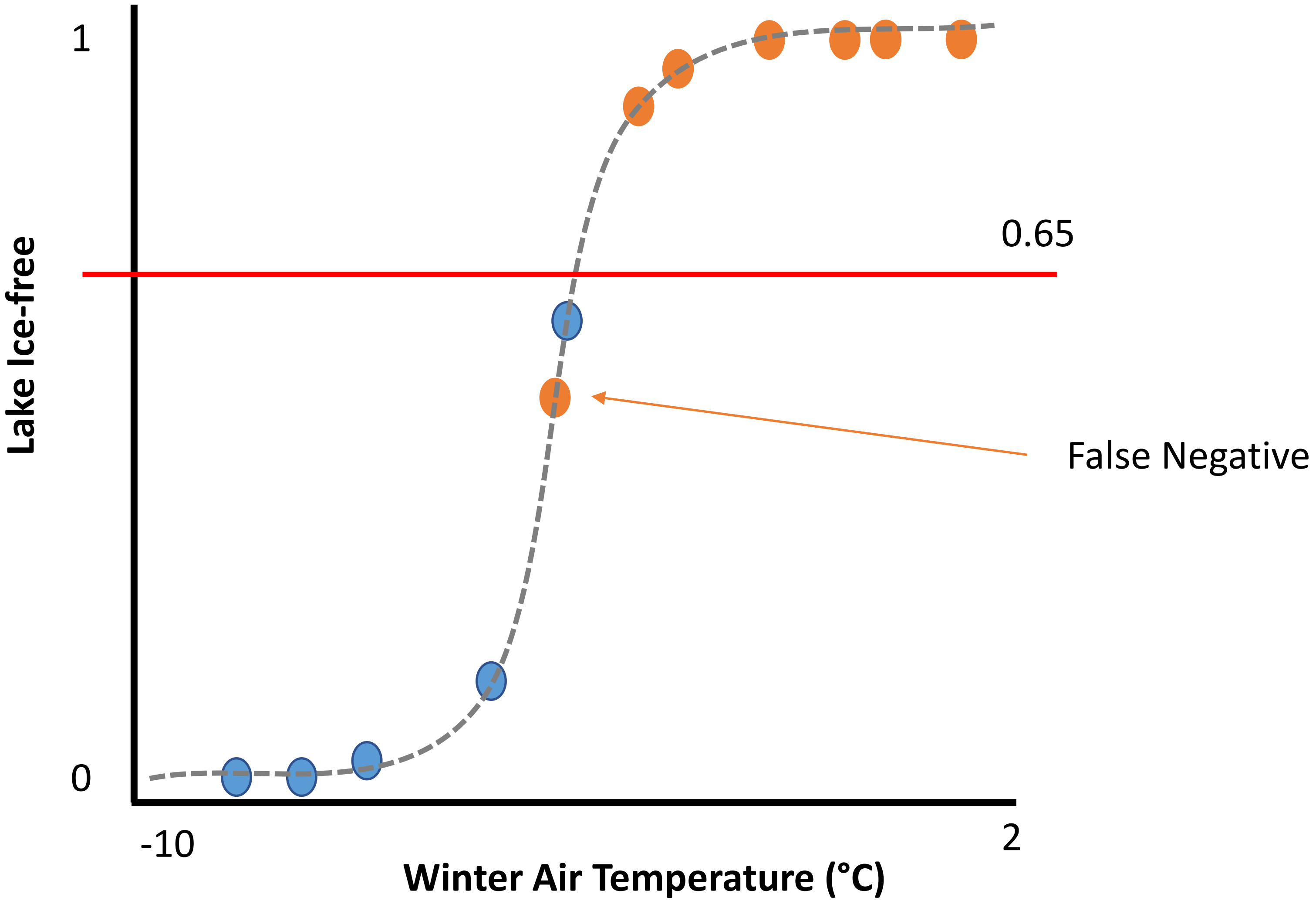
4) Prediction



4) Prediction



4) Prediction

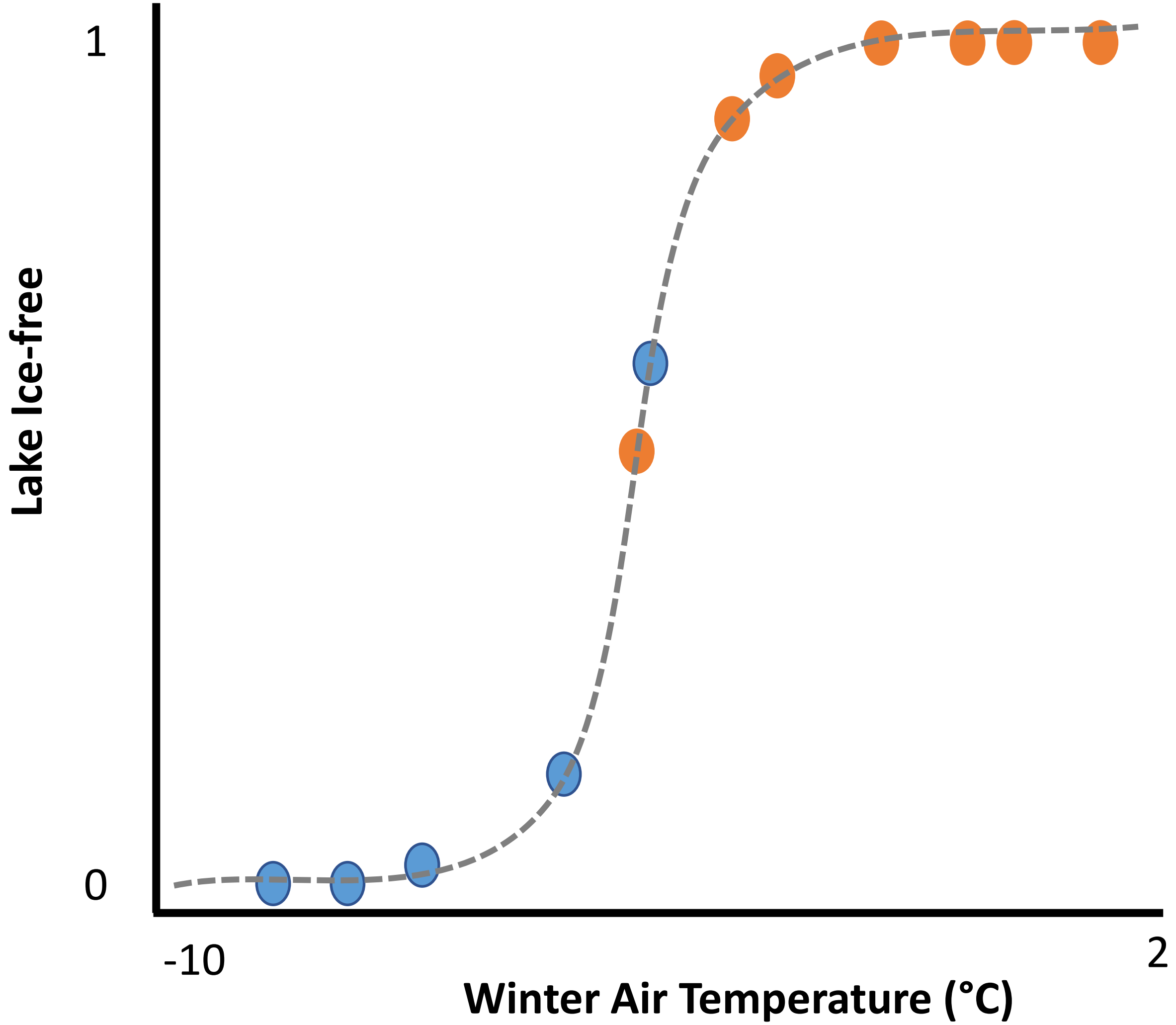


Question

When is an example where you would like to reduce false negative rate at the expensive of false positive?

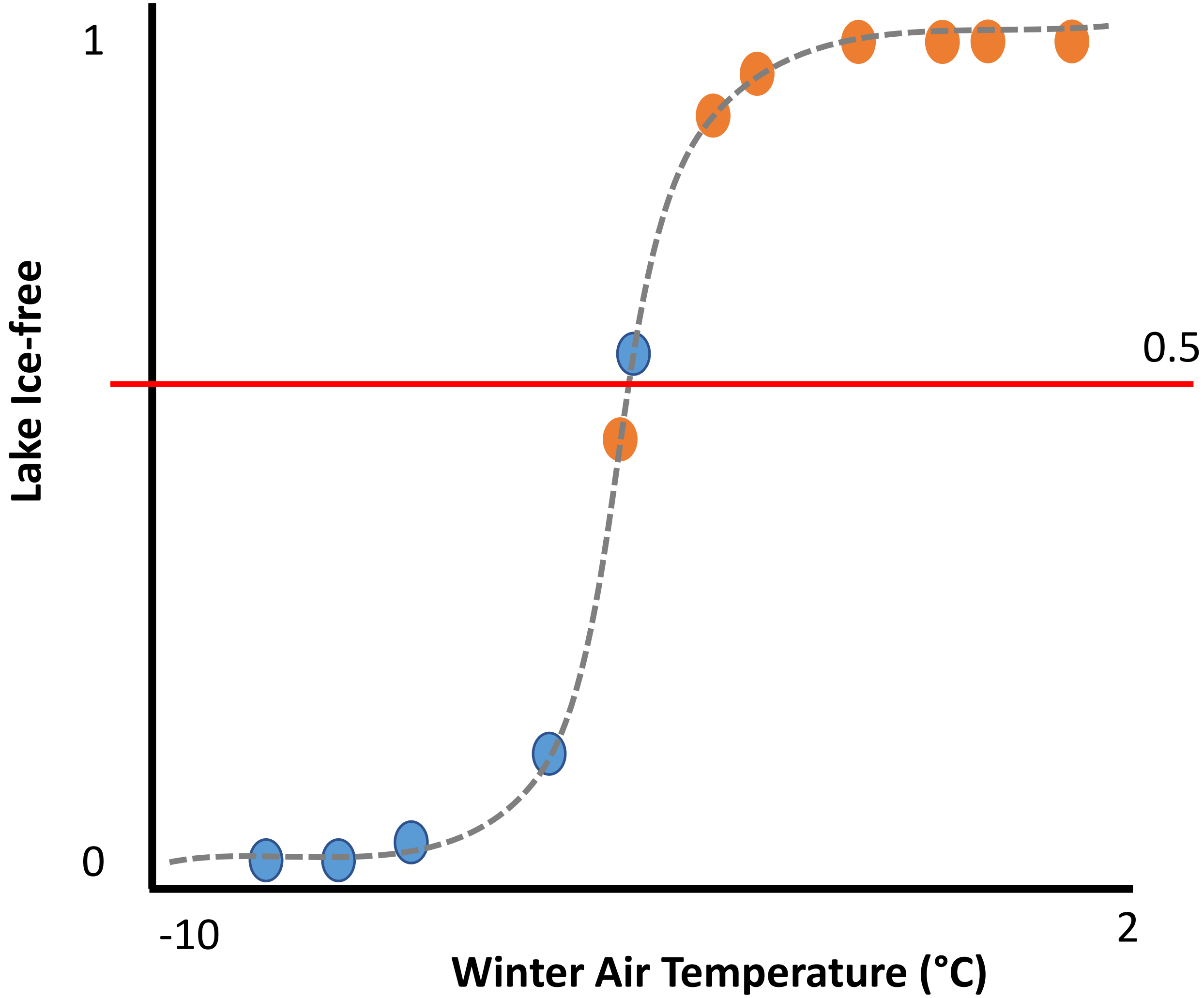
a) Predicting infection of a disease
b) Understanding climate effects on lake ice
c) Predicting the probability of a species occurring
d) A and C
e) All the above

4) Prediction – confusion matrix



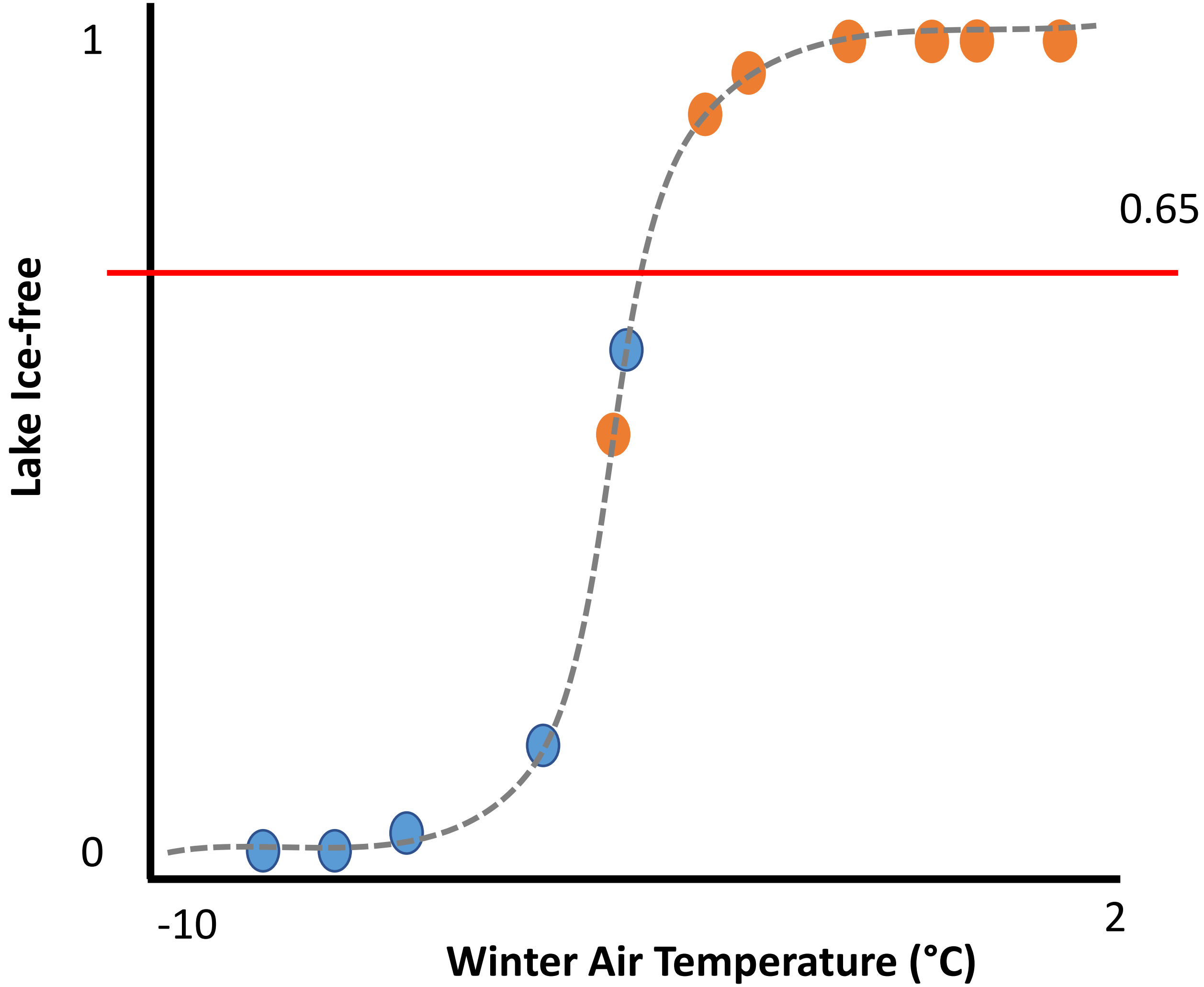
	Actual Positive	**Actual Negative**
Predicted Positive	True positive	False positive
Predicted Negative	False negative	True negative

4) Prediction – confusion matrix



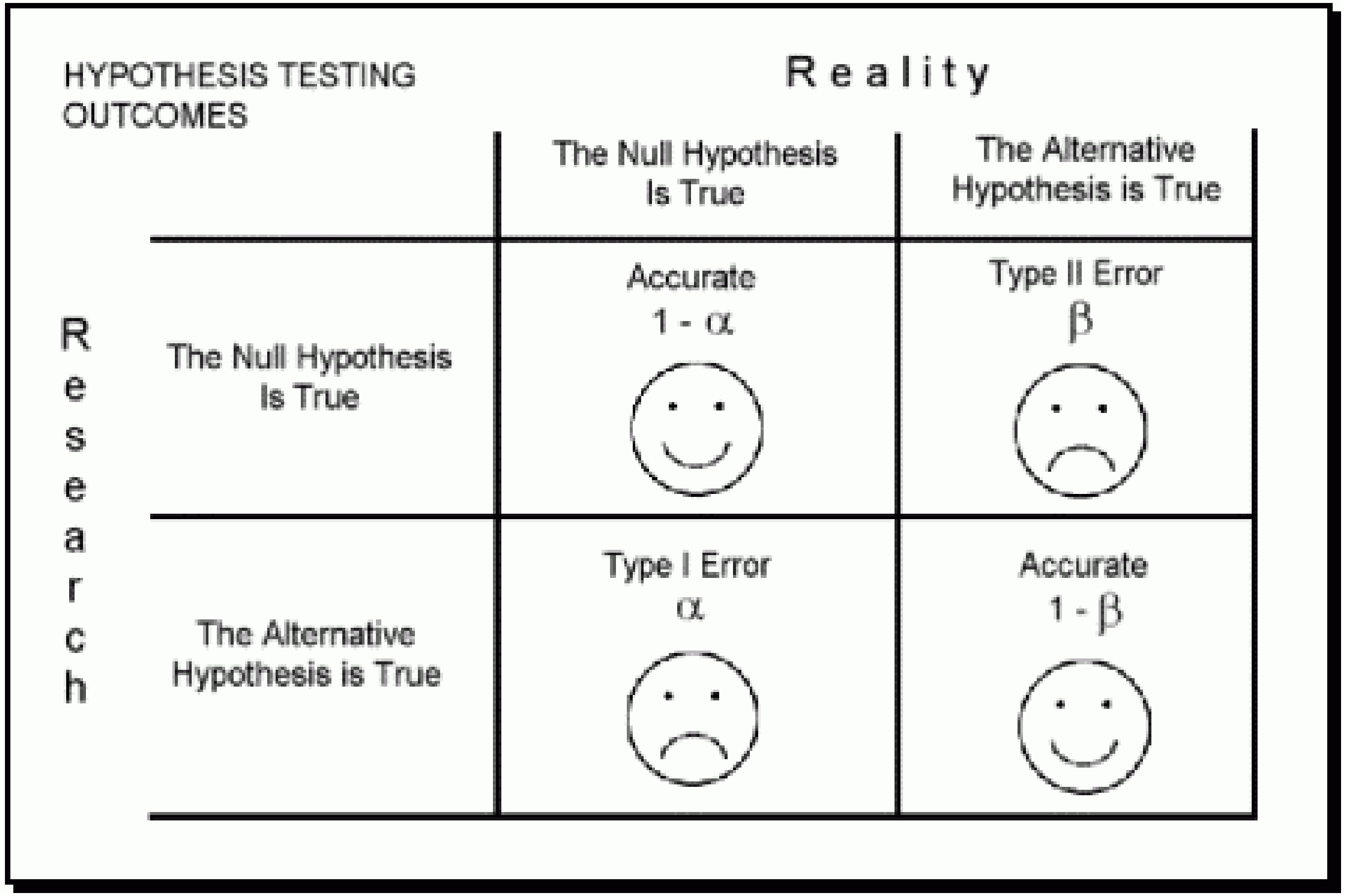
	Actual Positive	Actual Negative
Predicted Positive		
Predicted Negative		

4) Prediction – confusion matrix



	Actual Positive	Actual Negative
Predicted Positive		
Predicted Negative		

4) Prediction – confusion matrix



HYPOTHESIS TESTING OUTCOMES		Reality	
		The Null Hypothesis Is True	The Alternative Hypothesis is True
Research	The Null Hypothesis Is True	Accurate $1 - \alpha$	Type II Error β
	The Alternative Hypothesis is True	Type I Error α	Accurate $1 - \beta$

https://blog.datasciencedojo.com/

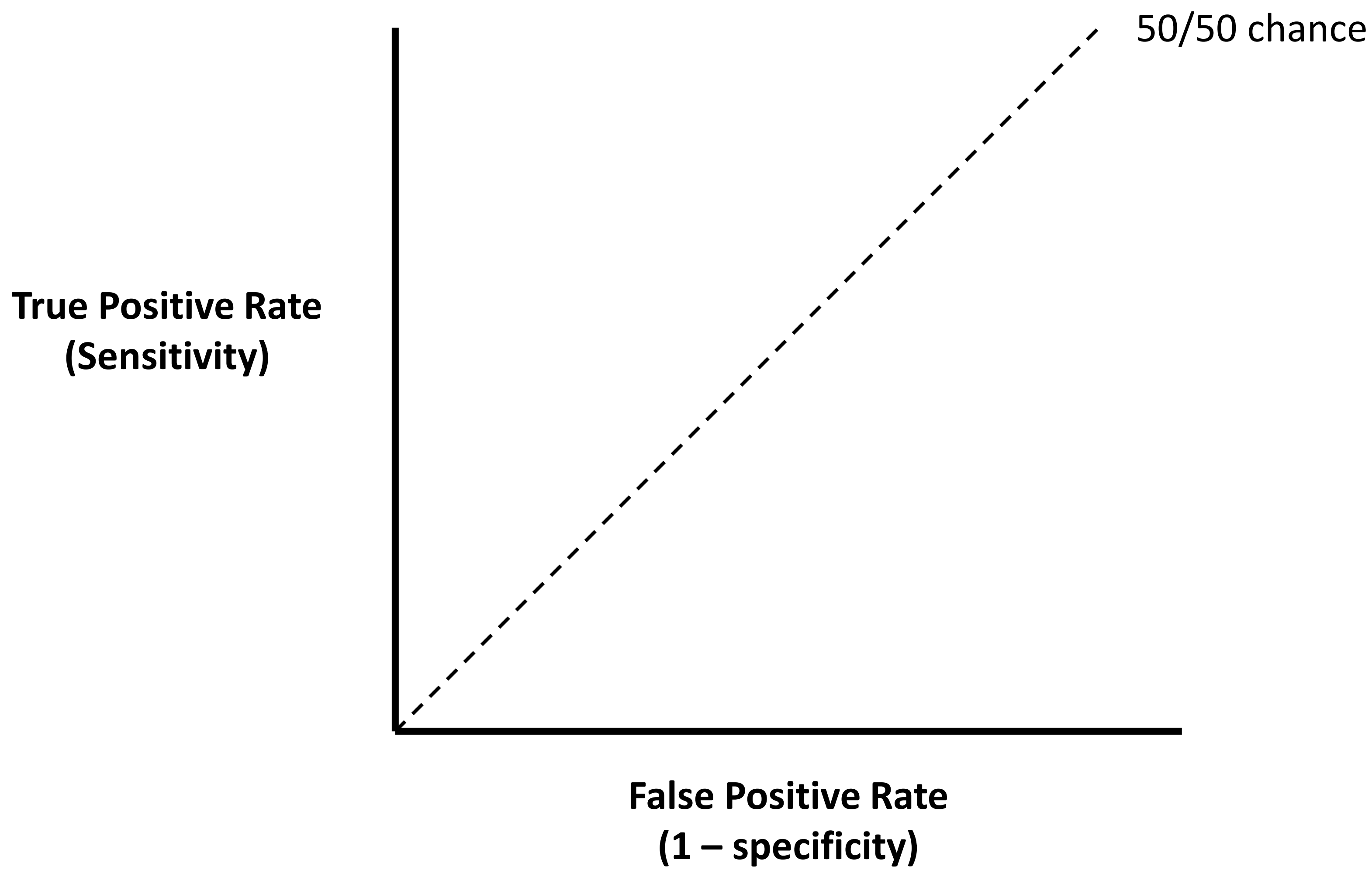
4) Optimizing the threshold

Many different methods

Receiver operator characteristic (ROC) often used to pick threshold

Area under the curve (AUC) of ROC used to compare models

4) Optimizing the threshold - ROC



4) Optimizing the threshold - ROC

	Actual Positive	**Actual Negative**
Predicted Positive	True positive	False positive
Predicted Negative	False negative	True negative

$$\text{True positive rate} = \frac{\text{True positives}}{\text{Actual positives}}$$

$$\text{True positive rate} = \frac{\text{True positives}}{\text{True positives + False negatives}}$$

Sensitivity

4) Optimizing the threshold - ROC

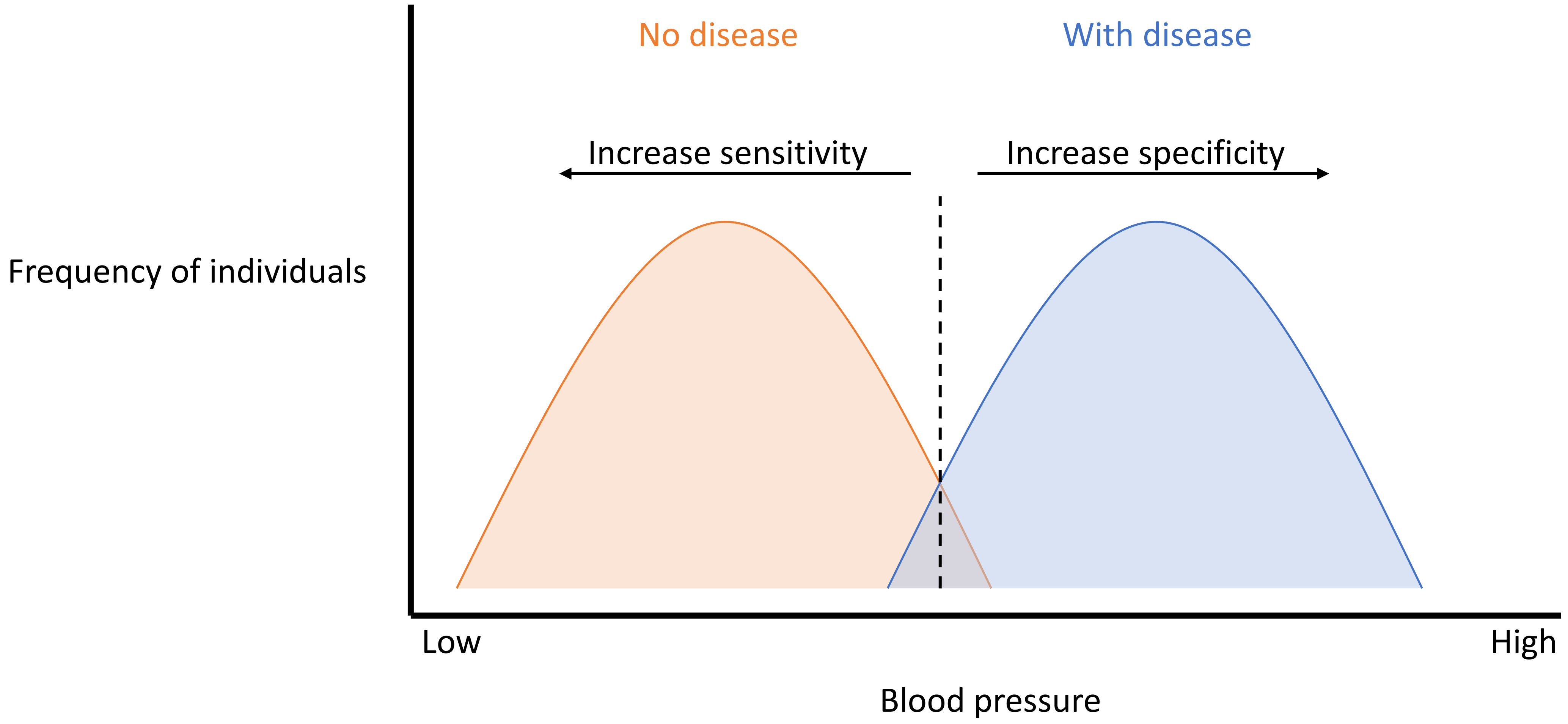
	Actual Positive	**Actual Negative**
Predicted Positive	True positive	False positive
Predicted Negative	False negative	True negative

$$\text{False positive rate} = \frac{\text{False positives}}{\text{Actual negatives}}$$

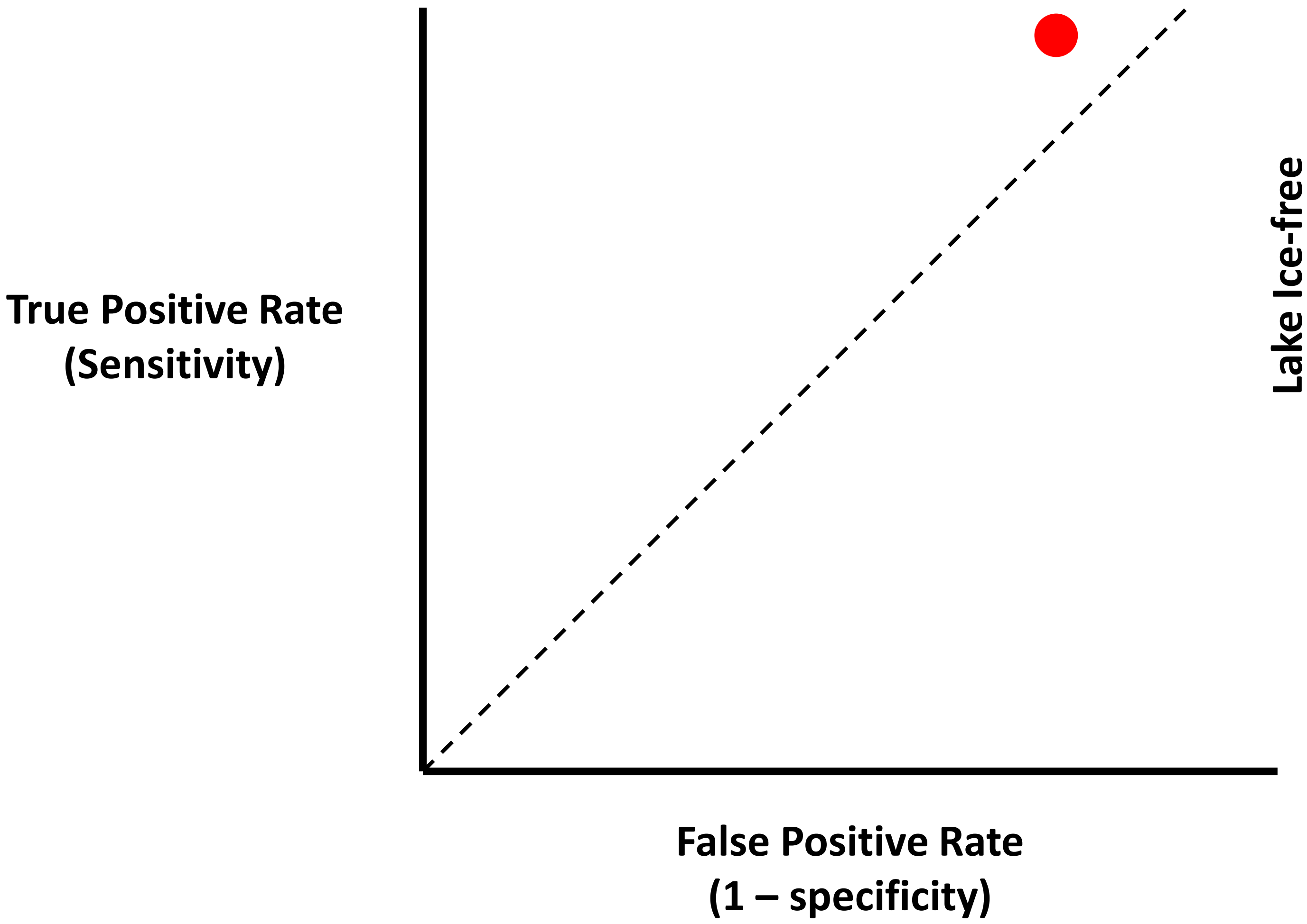
$$\text{False positive rate} = \frac{\text{True positives}}{\text{False positives + True negatives}}$$

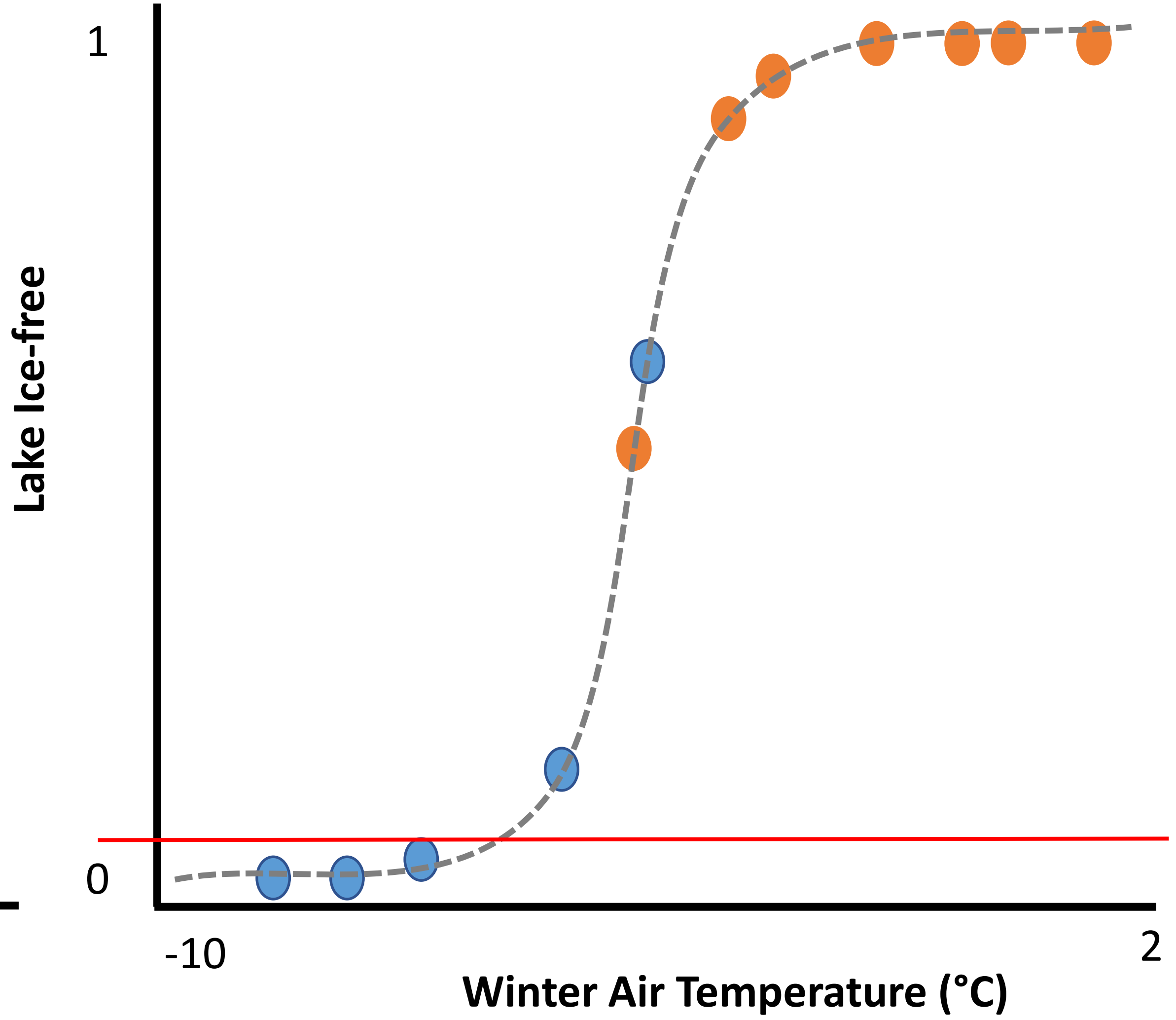
Specificity

4) Sensitivity vs. specificity

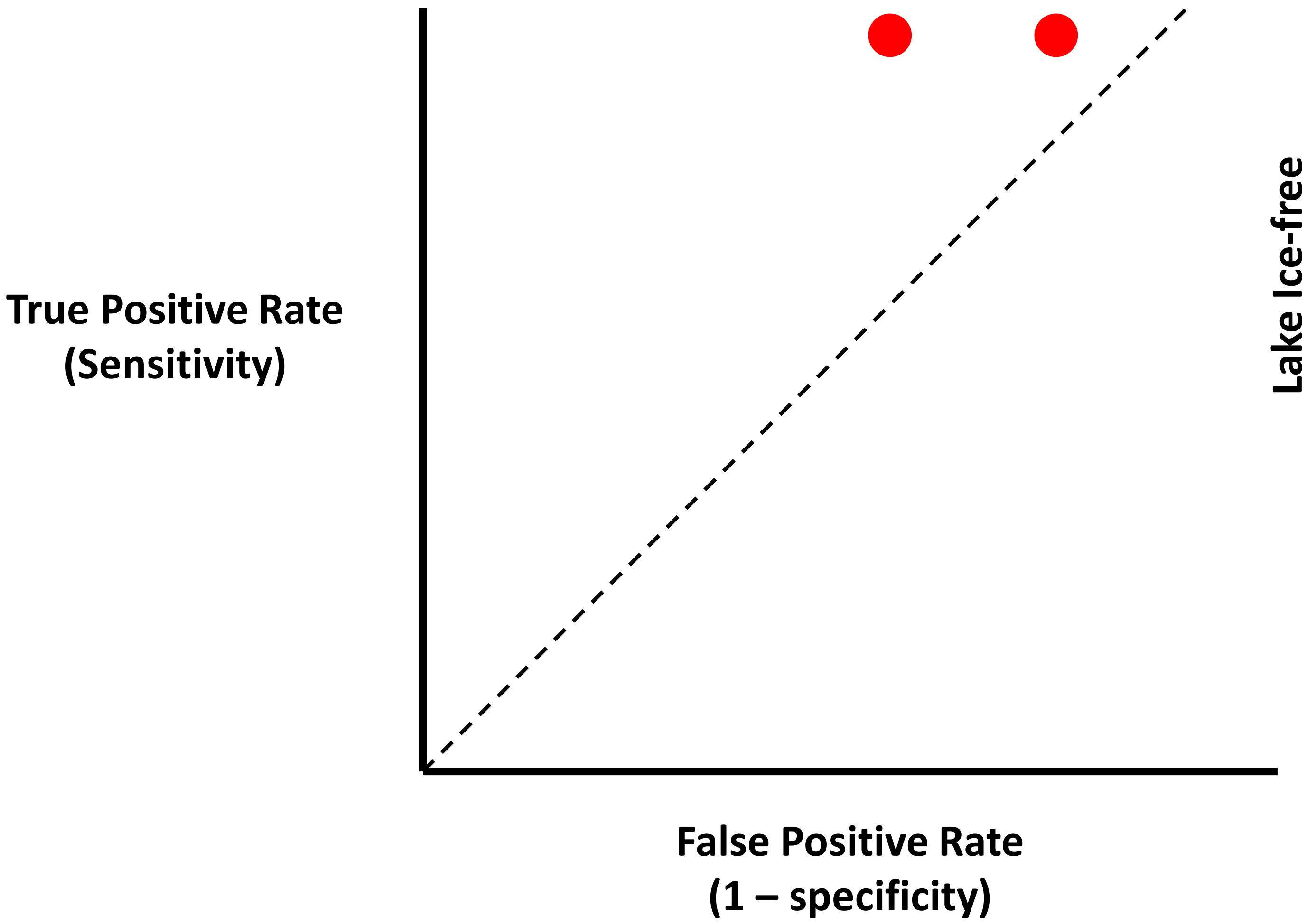


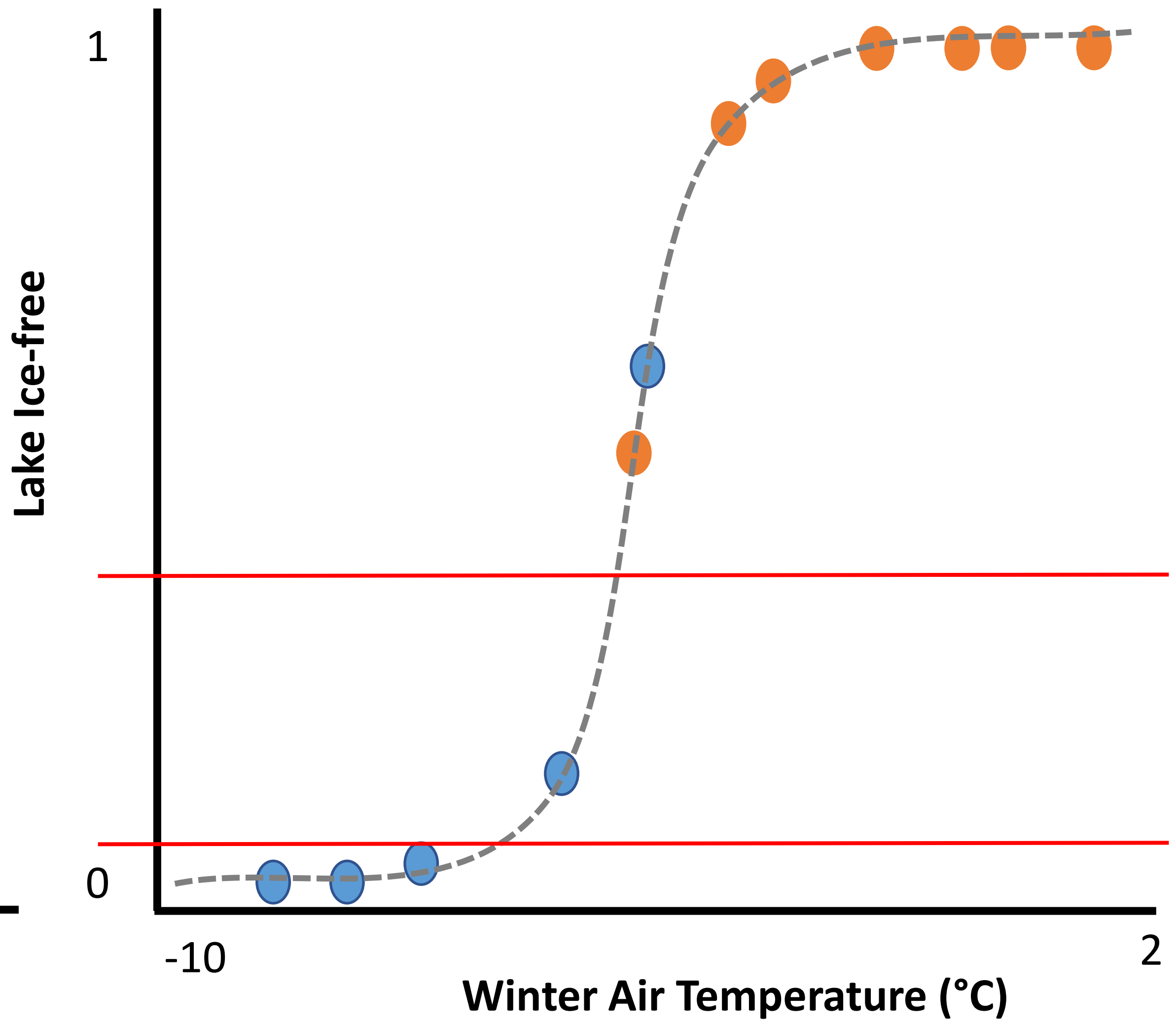
4) Optimizing the threshold - ROC



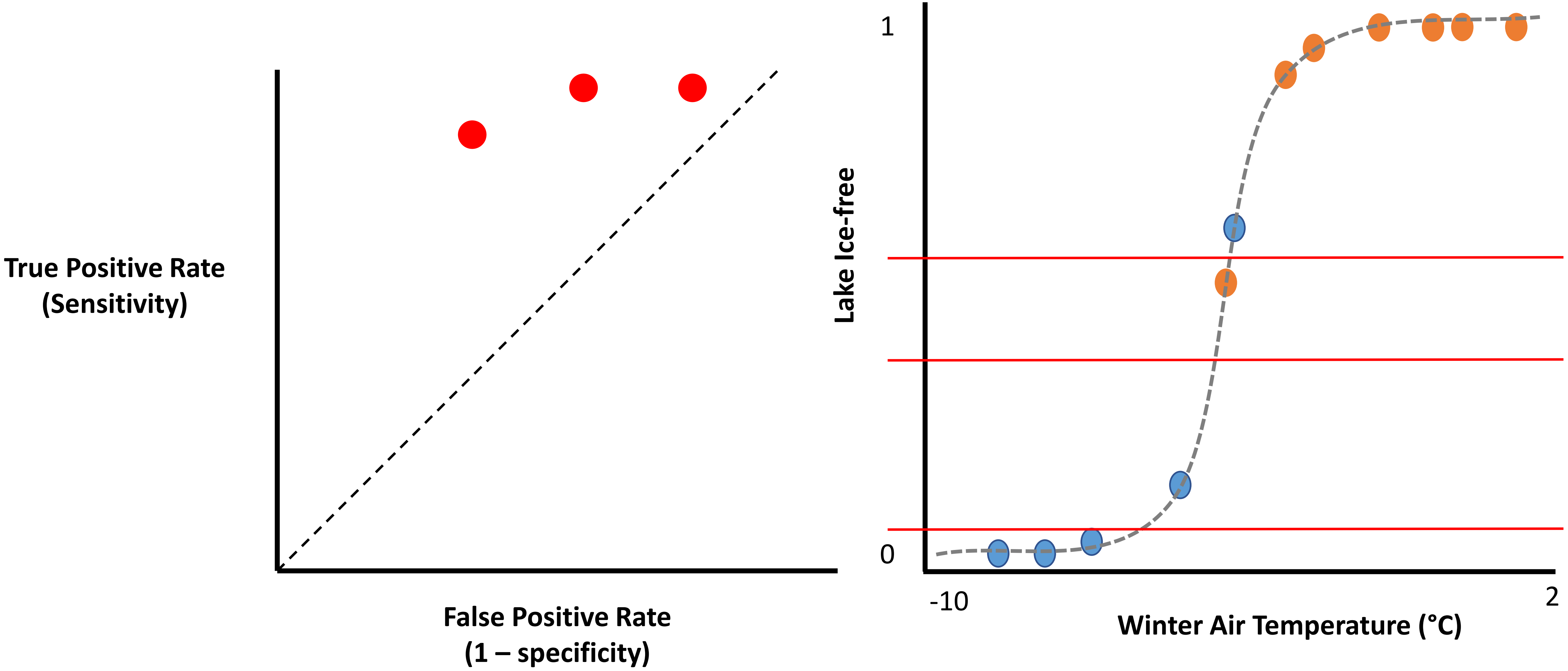


4) Optimizing the threshold - ROC

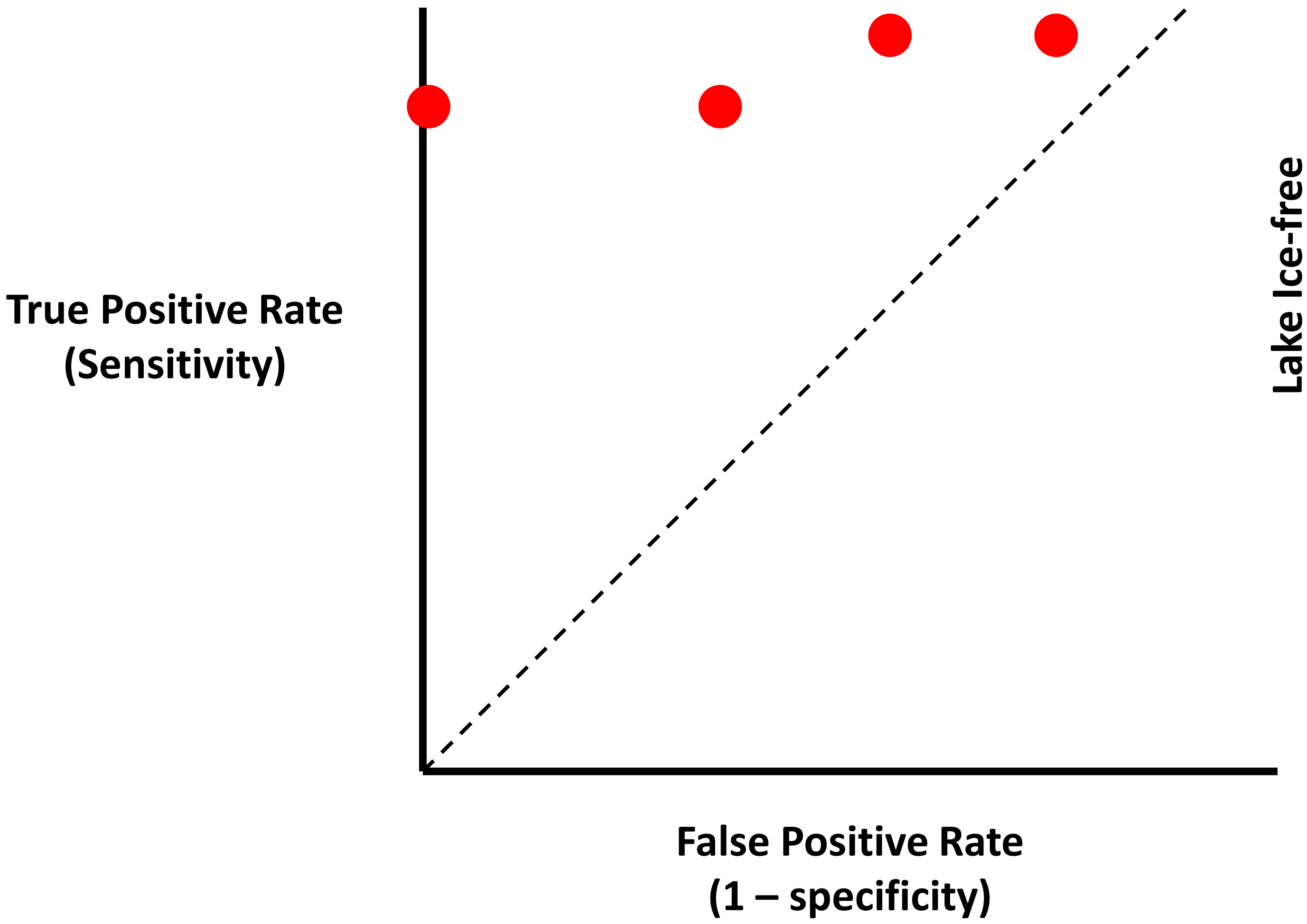


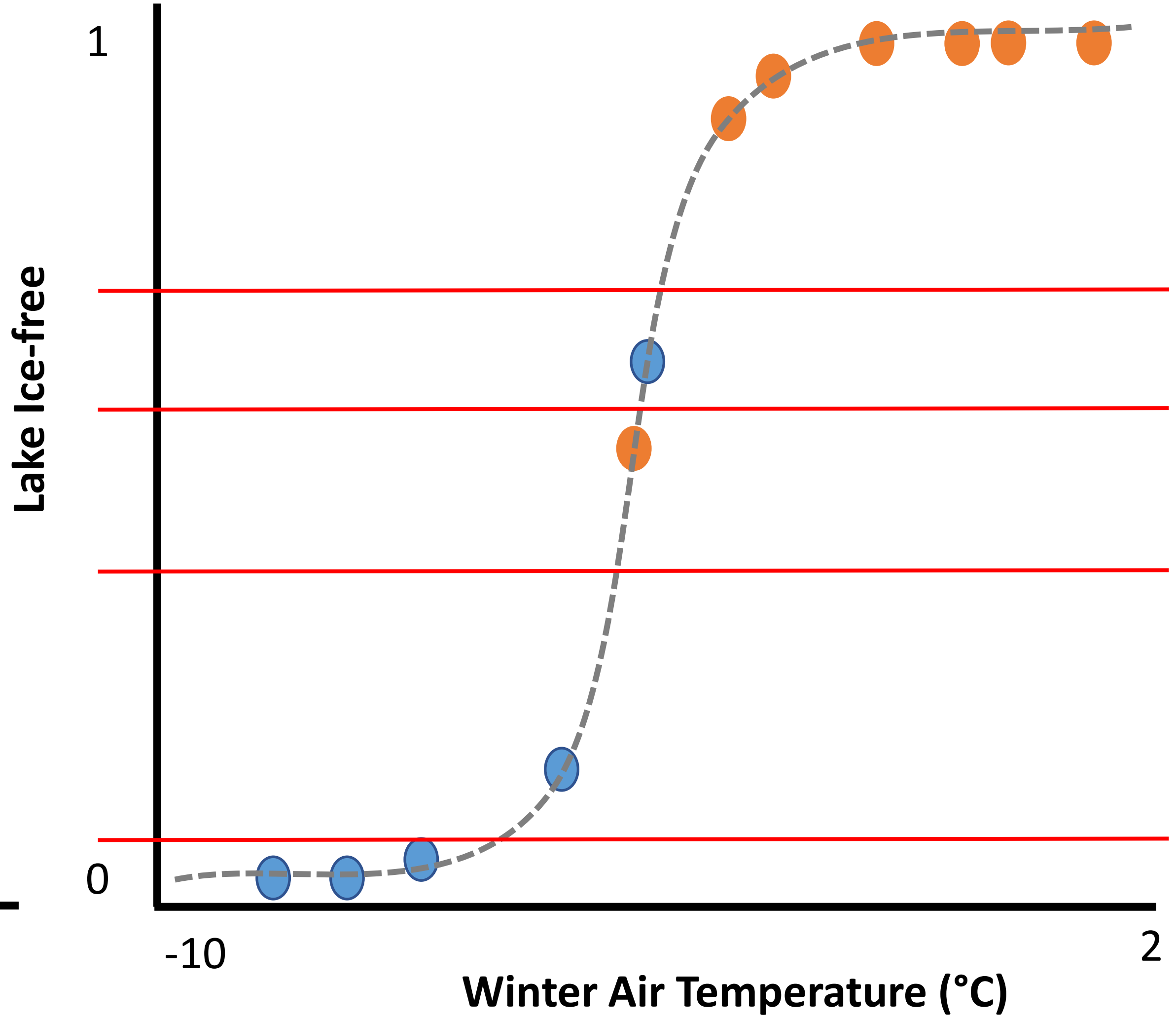


4) Optimizing the threshold - ROC

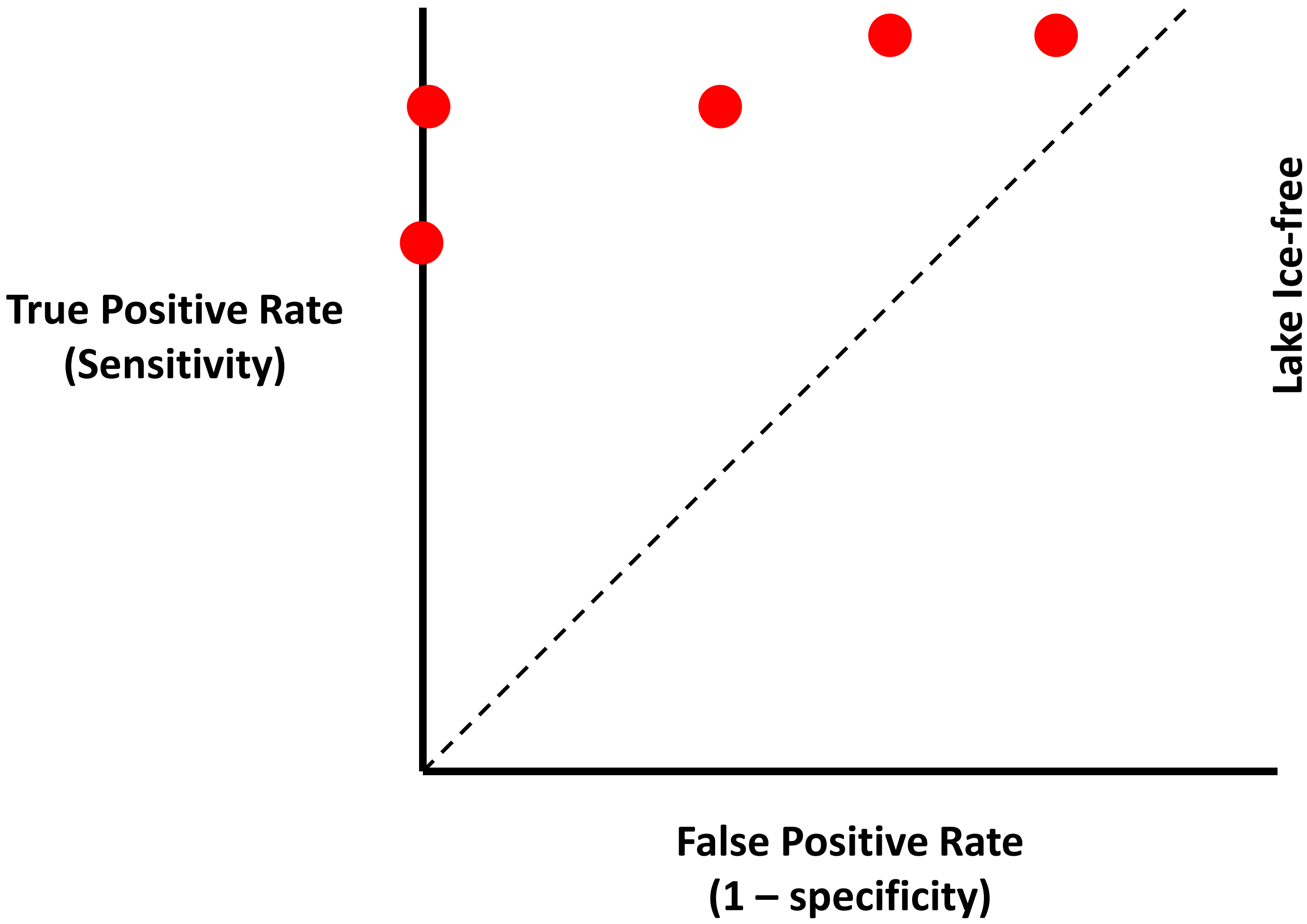


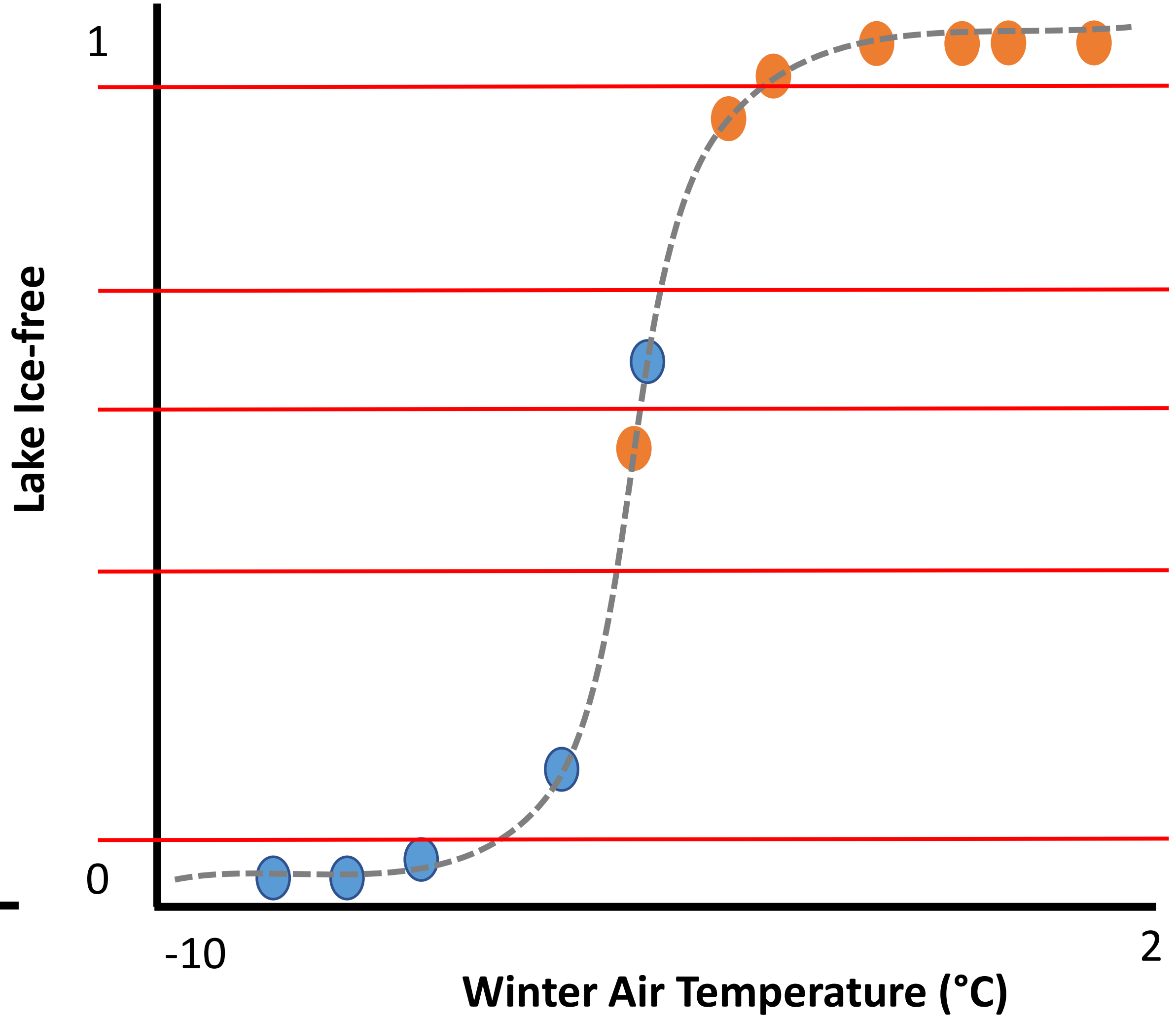
4) Optimizing the threshold - ROC



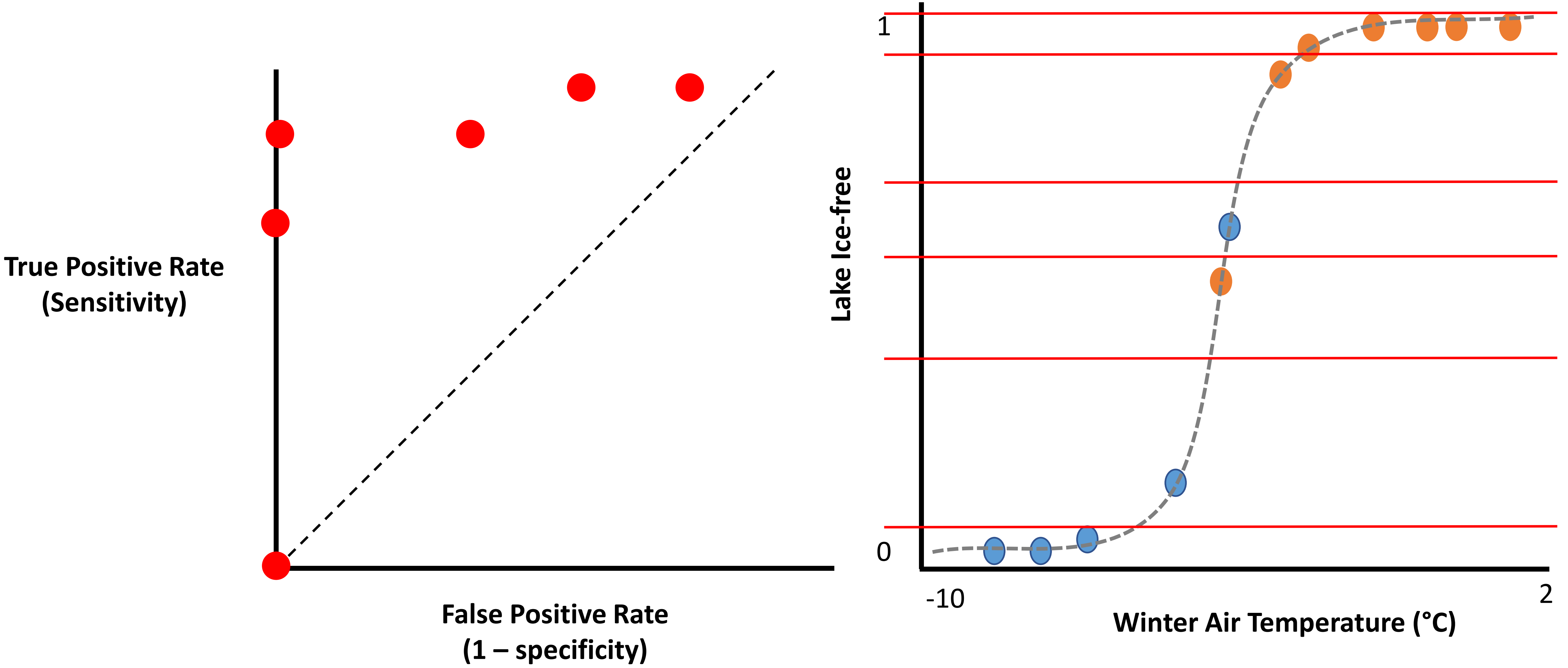


4) Optimizing the threshold - ROC



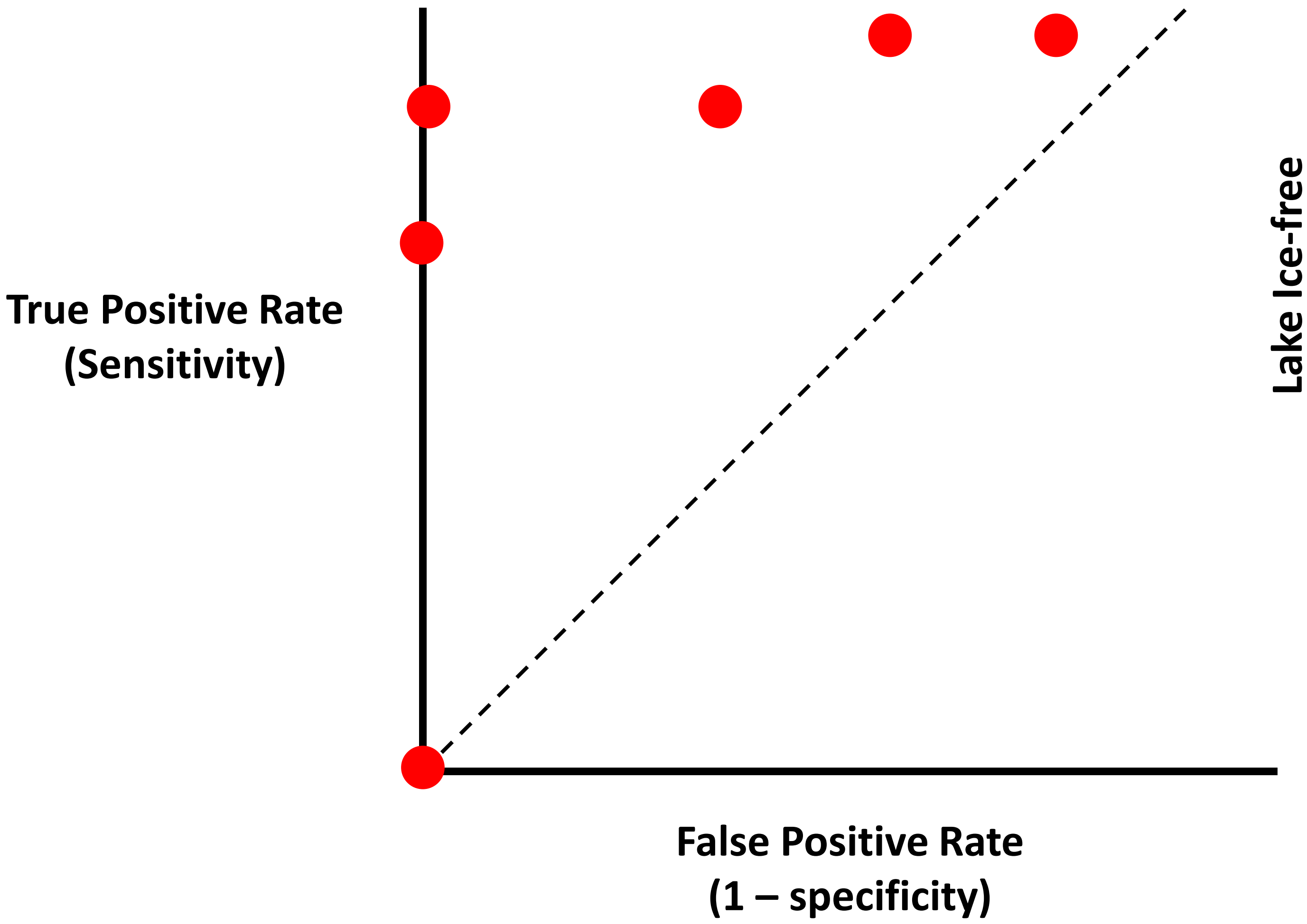


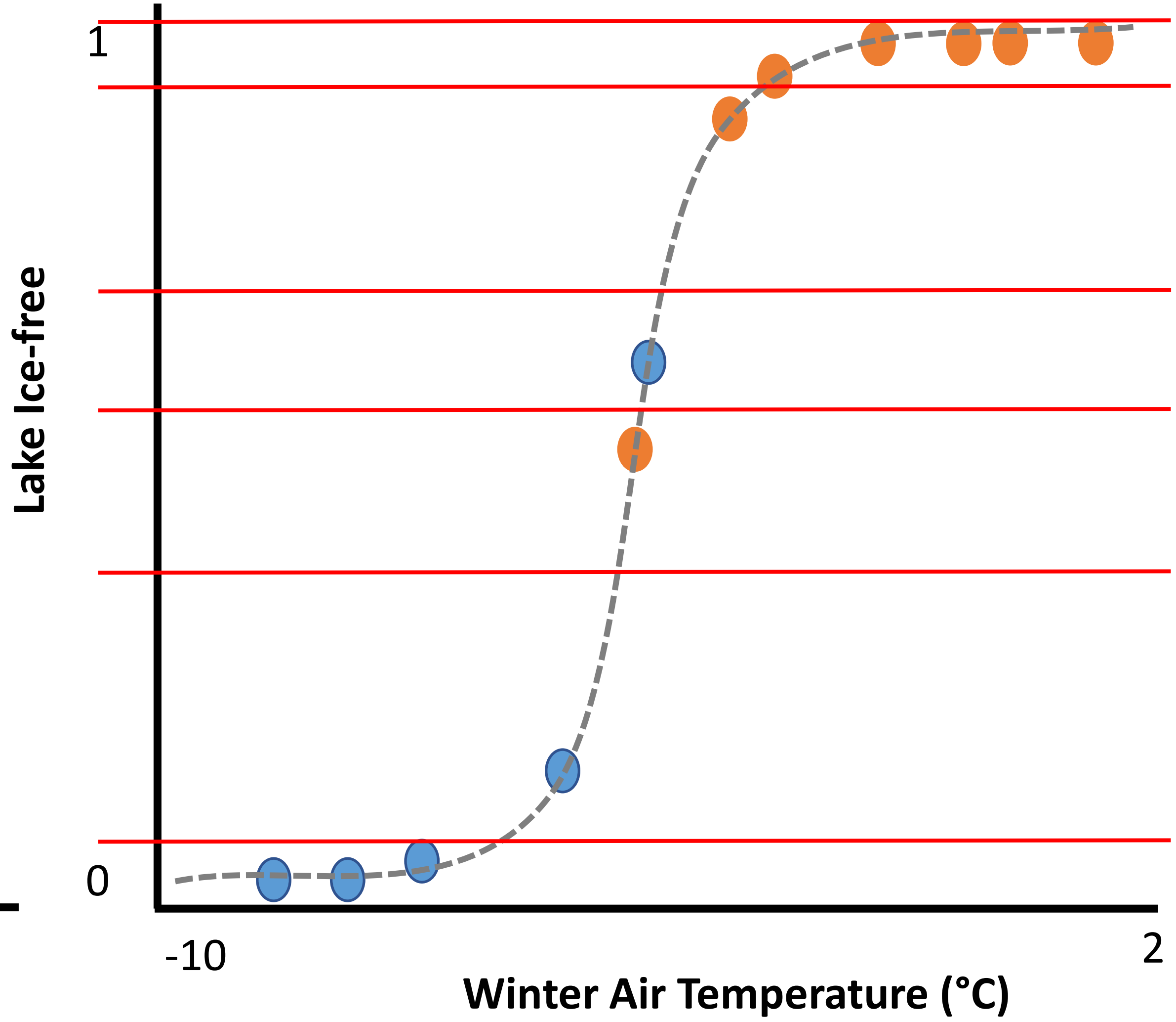
4) Optimizing the threshold - ROC



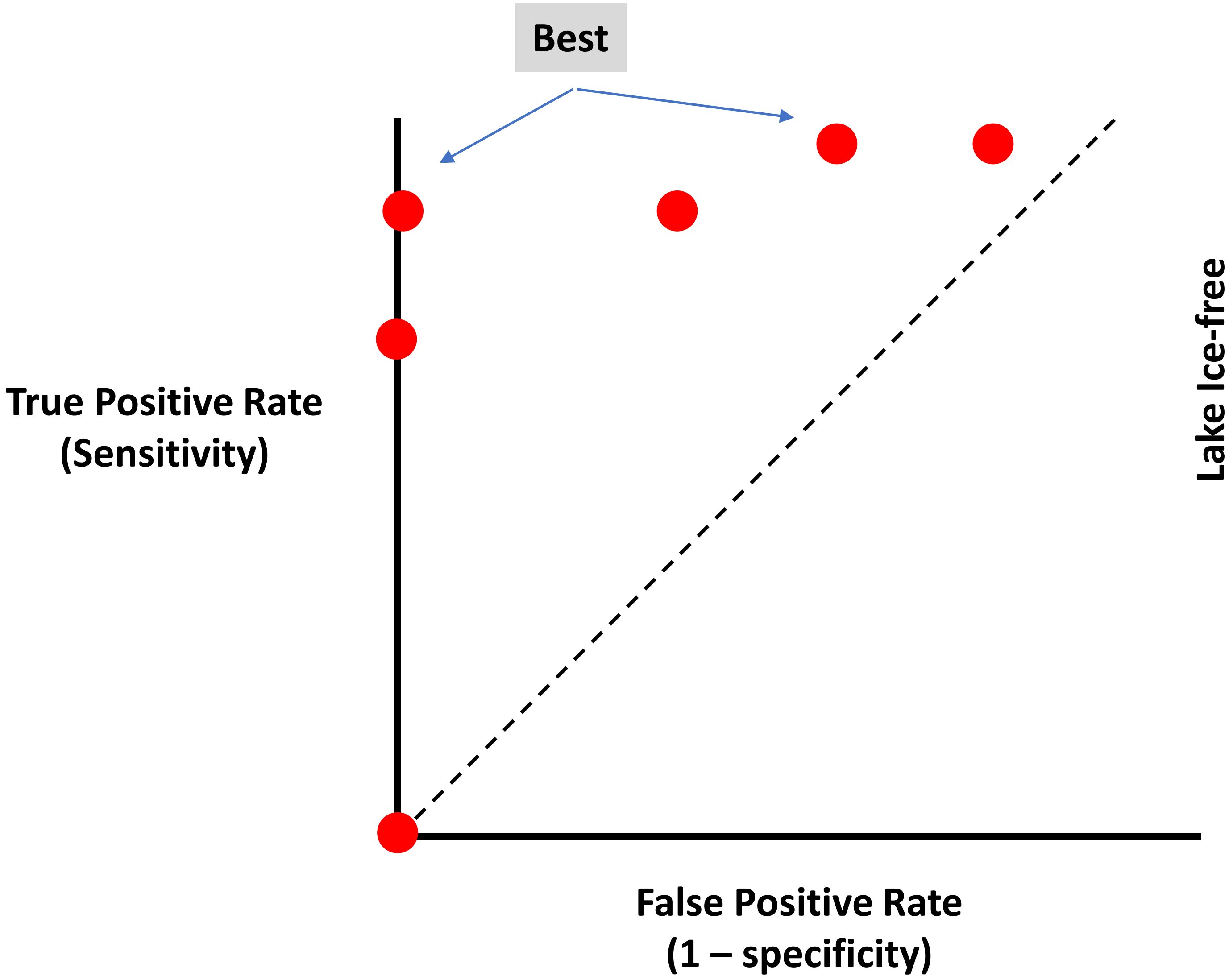
4) Optimizing the threshold - ROC

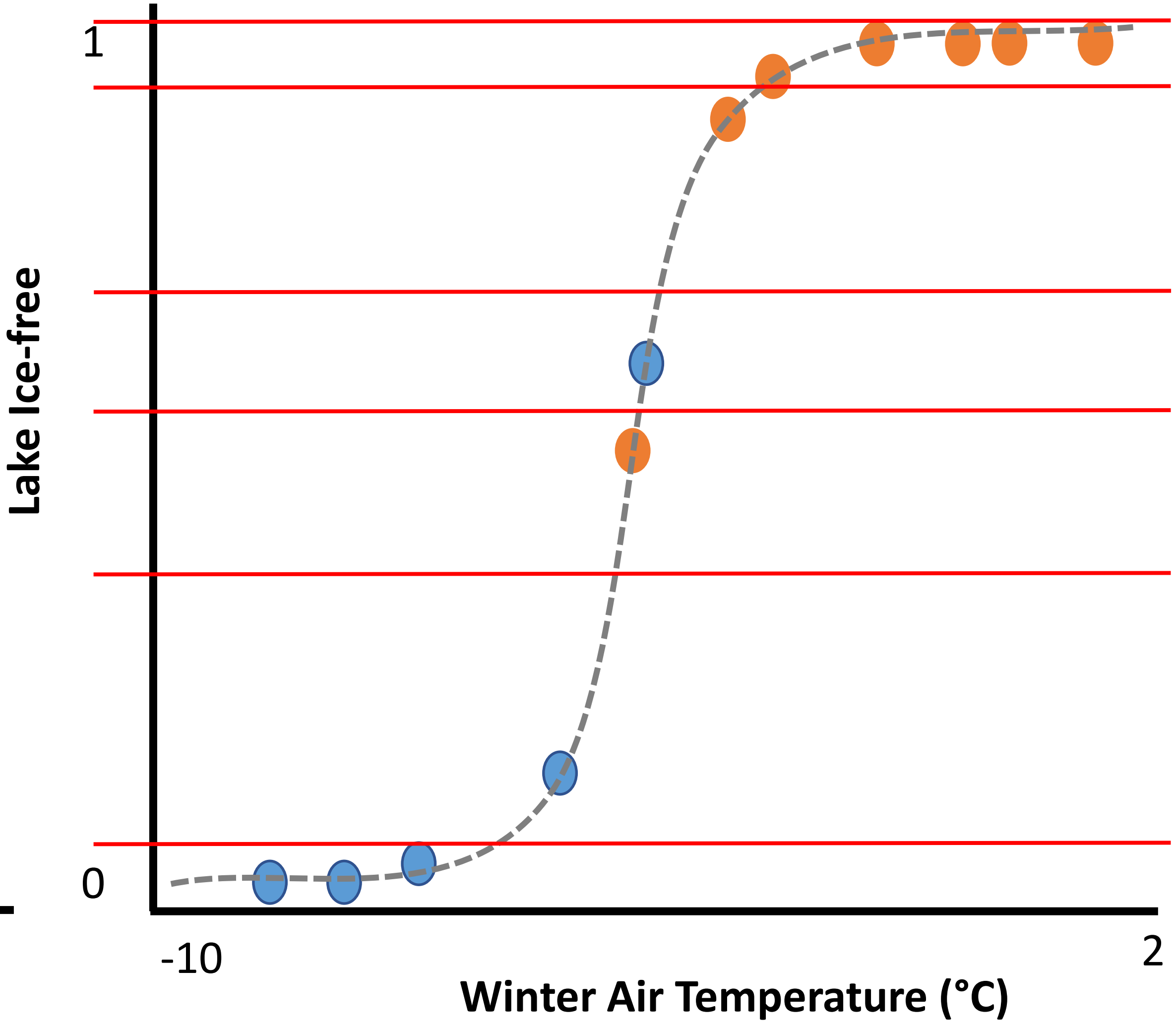
Q: Which is the best?



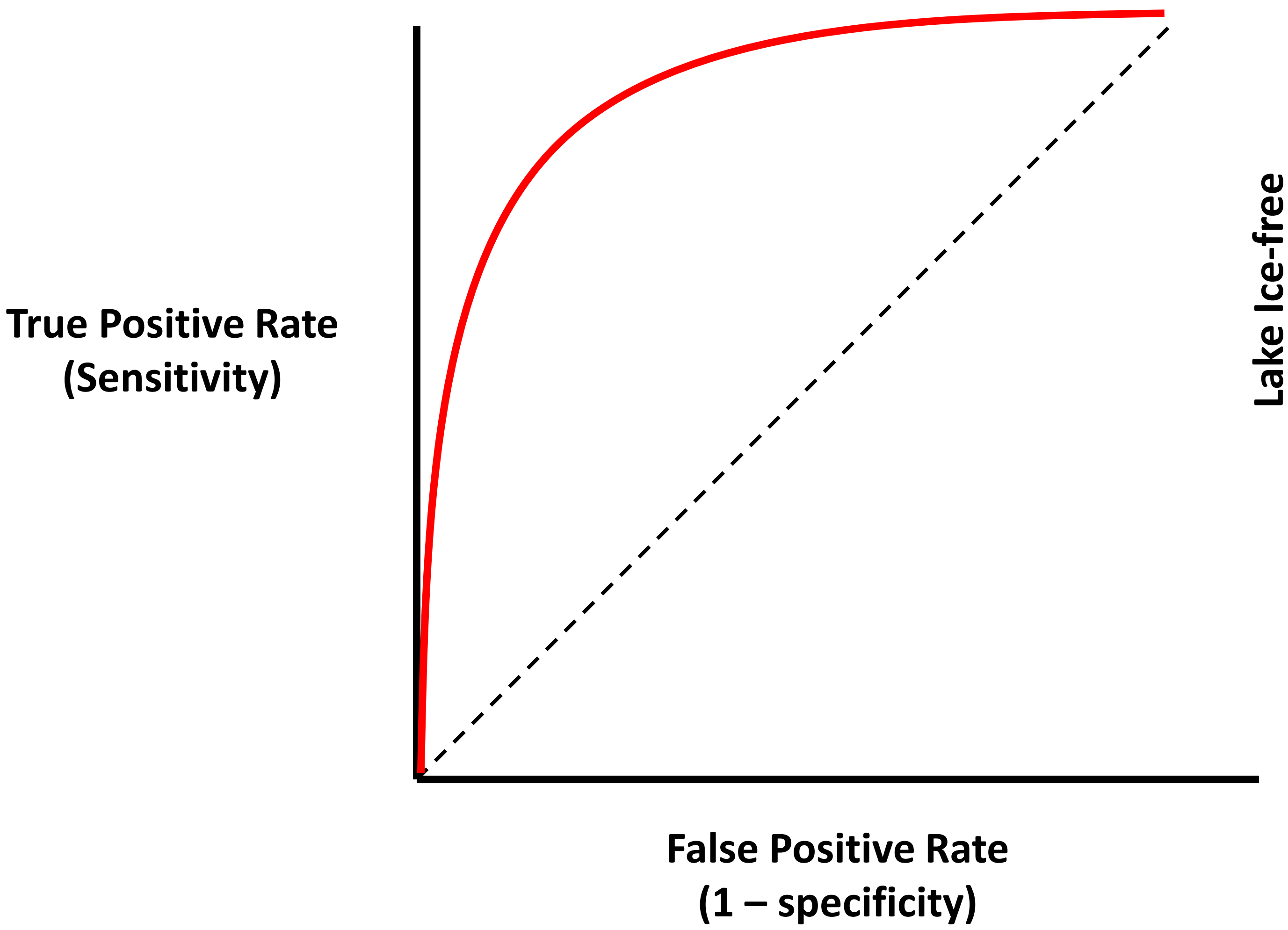


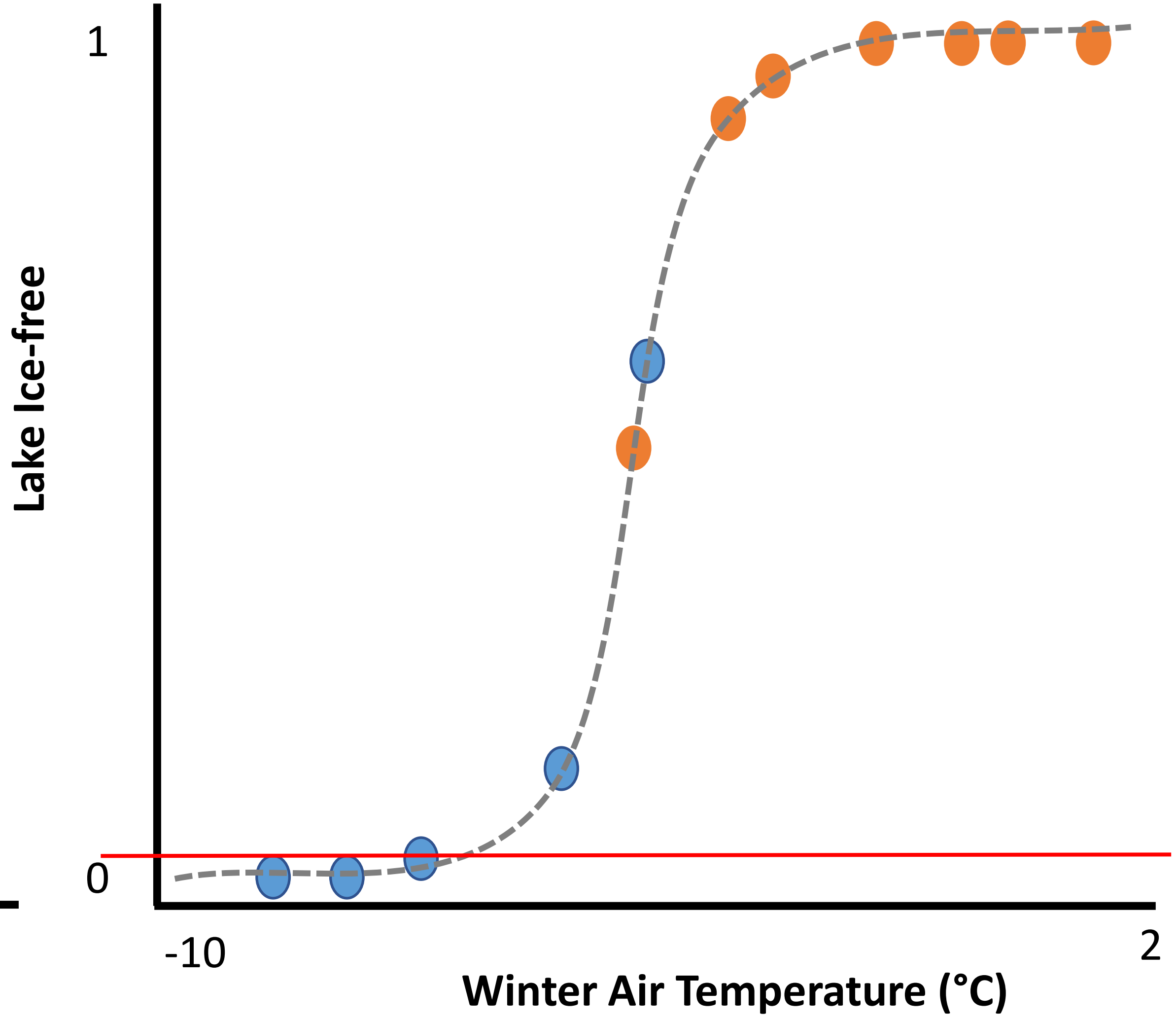
4) Optimizing the threshold - ROC



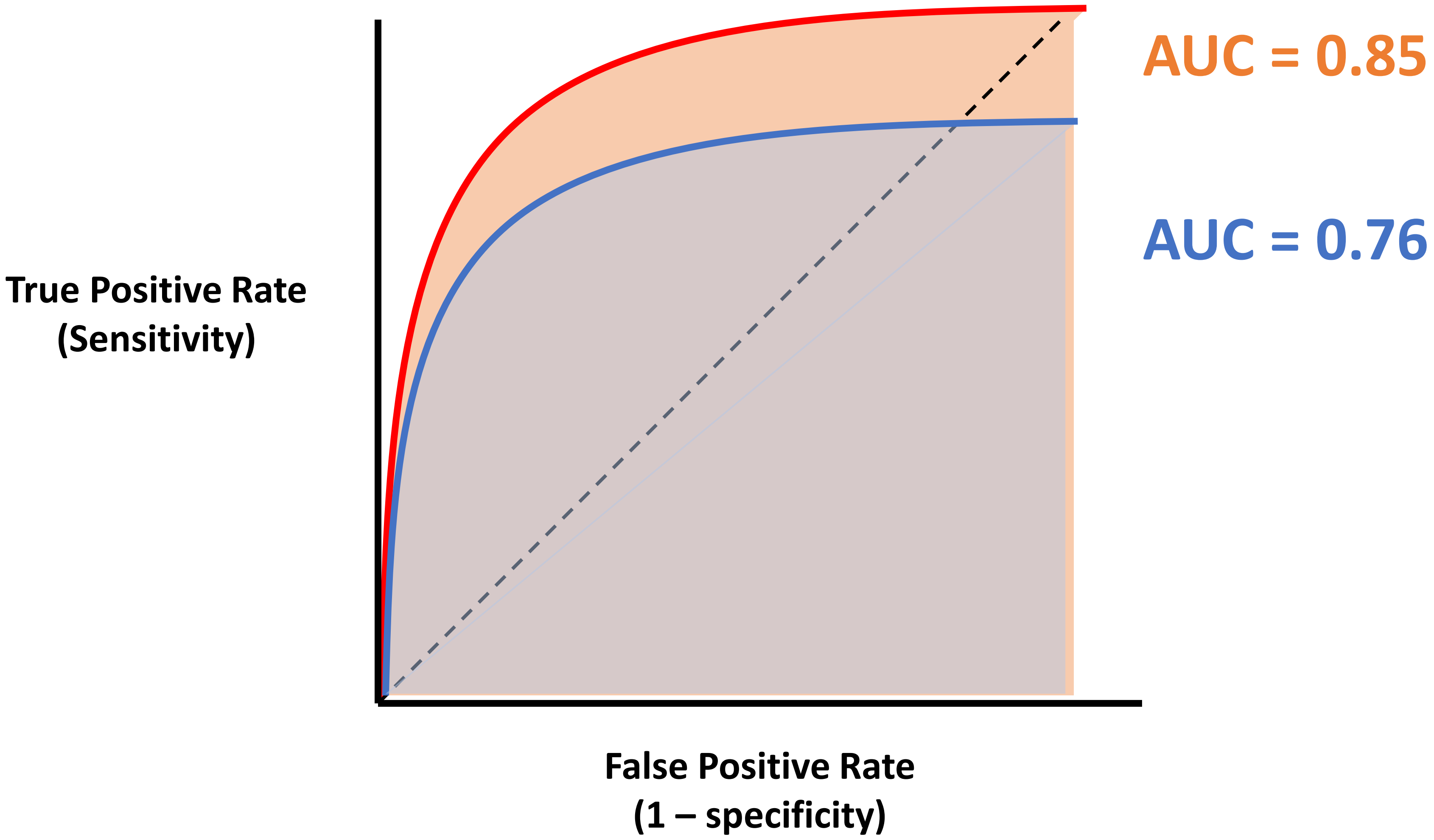


4) Optimizing the threshold - ROC





4) Comparing models - AUC

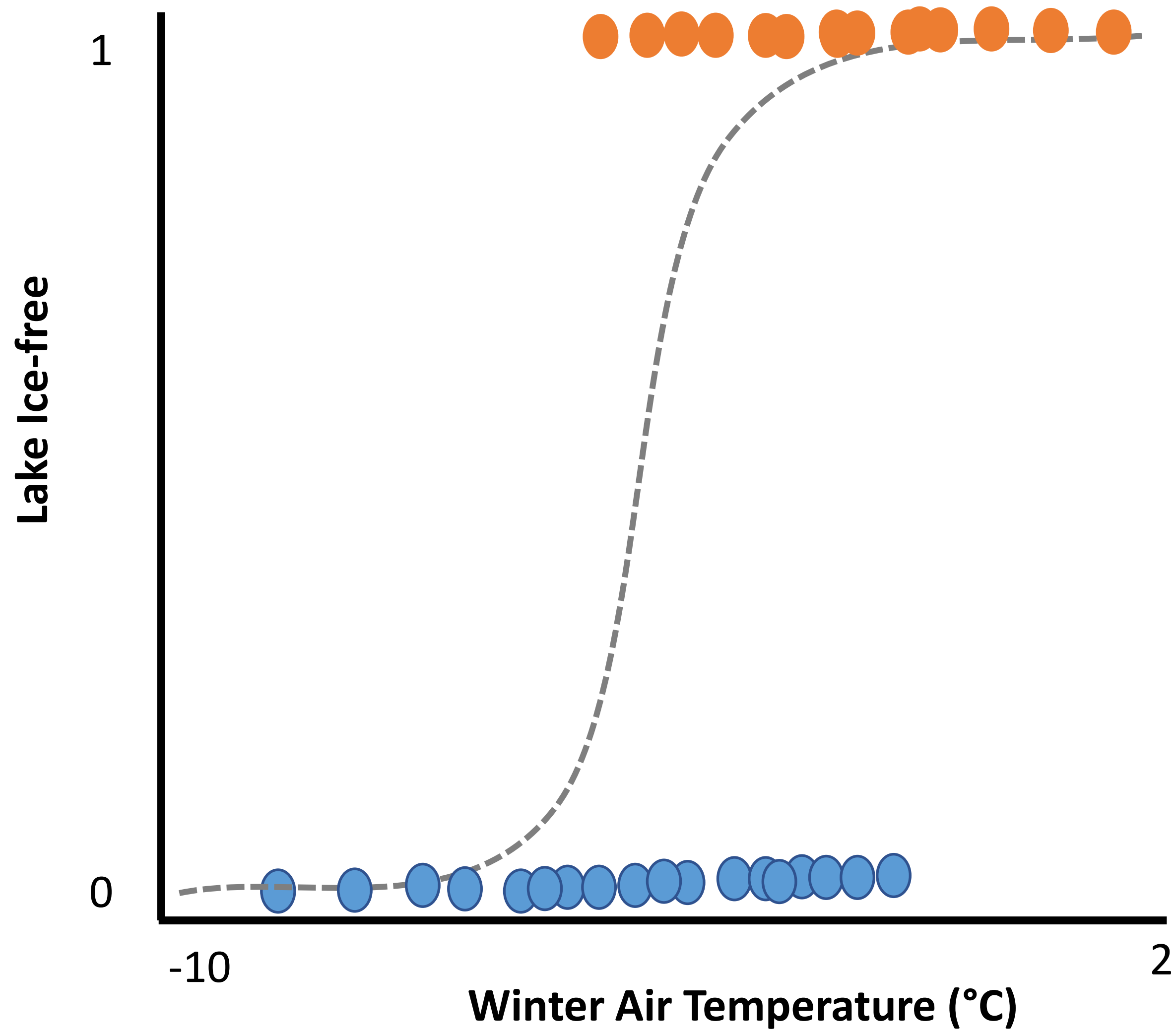


A note about overdispersion

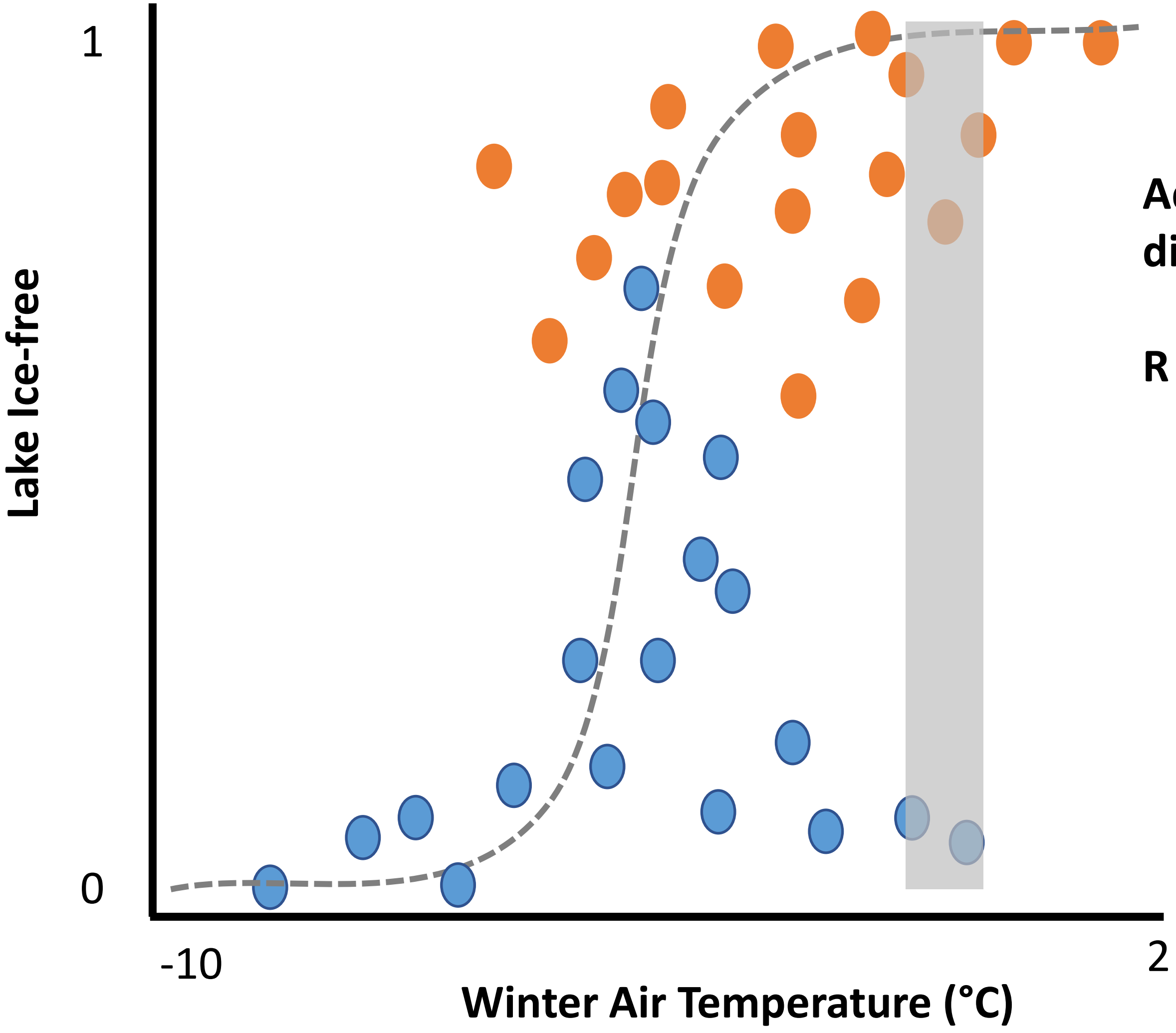
- Over dispersion when the variance exceeds the mean model fit

- Frequent in ecology data

A note about overdispersion



A note about overdispersion



Adjust for this using a dispersion parameter

R - Quasi-binomial

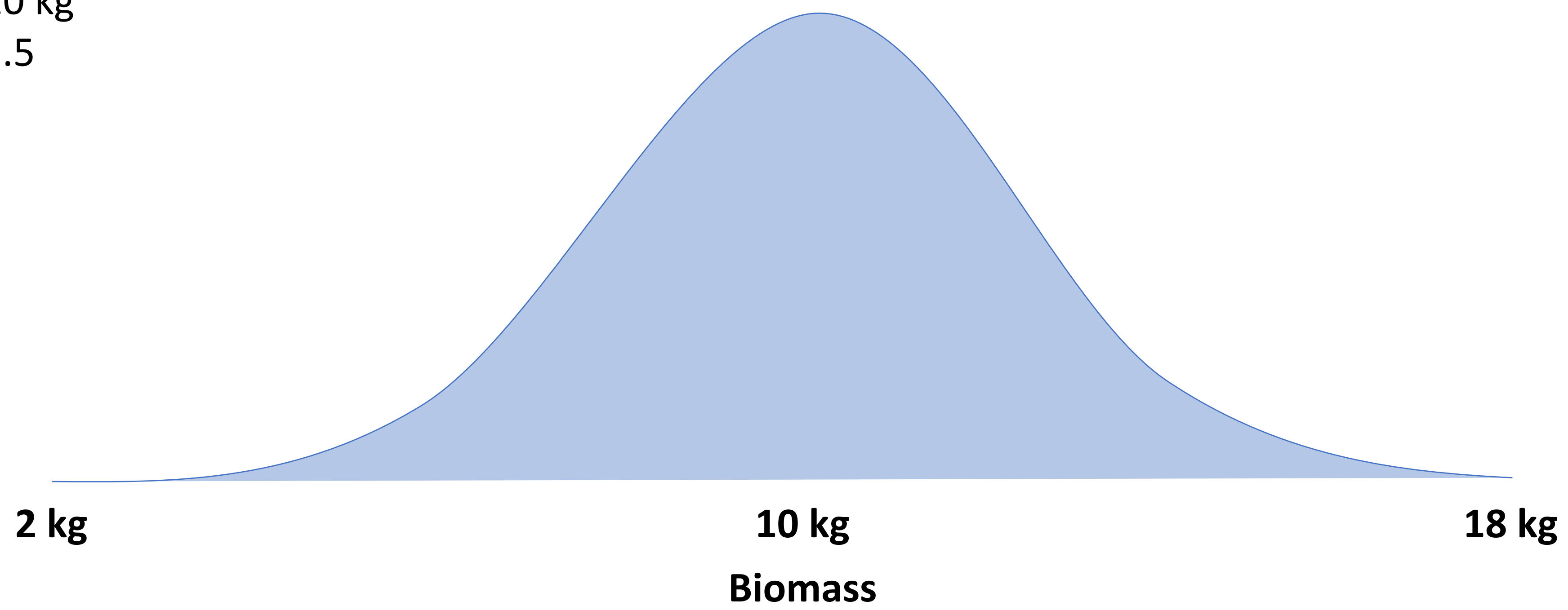
How do logistic regressions relate to GLMs?

- Use a link function to connect to a linear function
- Use Maximum Likelihood rather than sum of squares
- Allows for the flexibility in GLMs for many distributions

GLM using a normal distribution

Probability vs. Likelihood

Normal distribution
μ = 10 kg
σ = 3.5

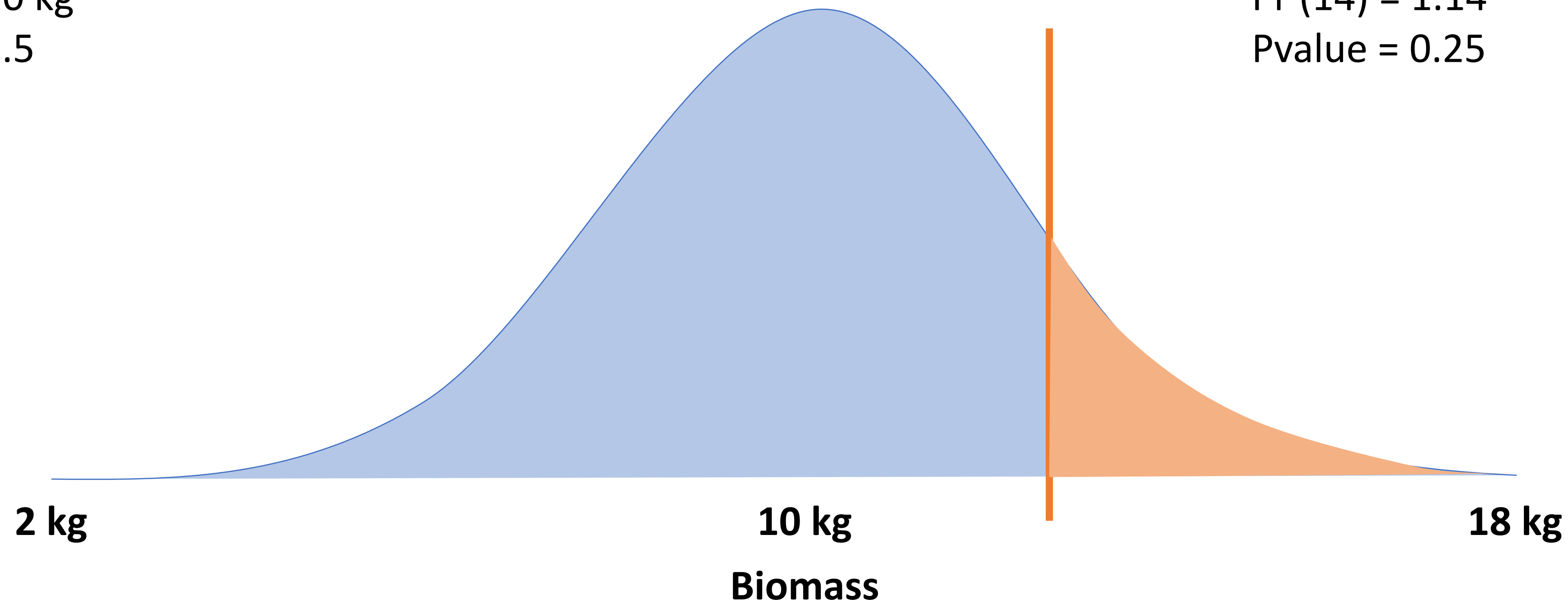


GLM using a normal distribution

Probability vs. Likelihood

Normal distribution
μ = 10 kg
σ = 3.5

Pr (14) = ?
Pr (14) = 1.14
Pvalue = 0.25

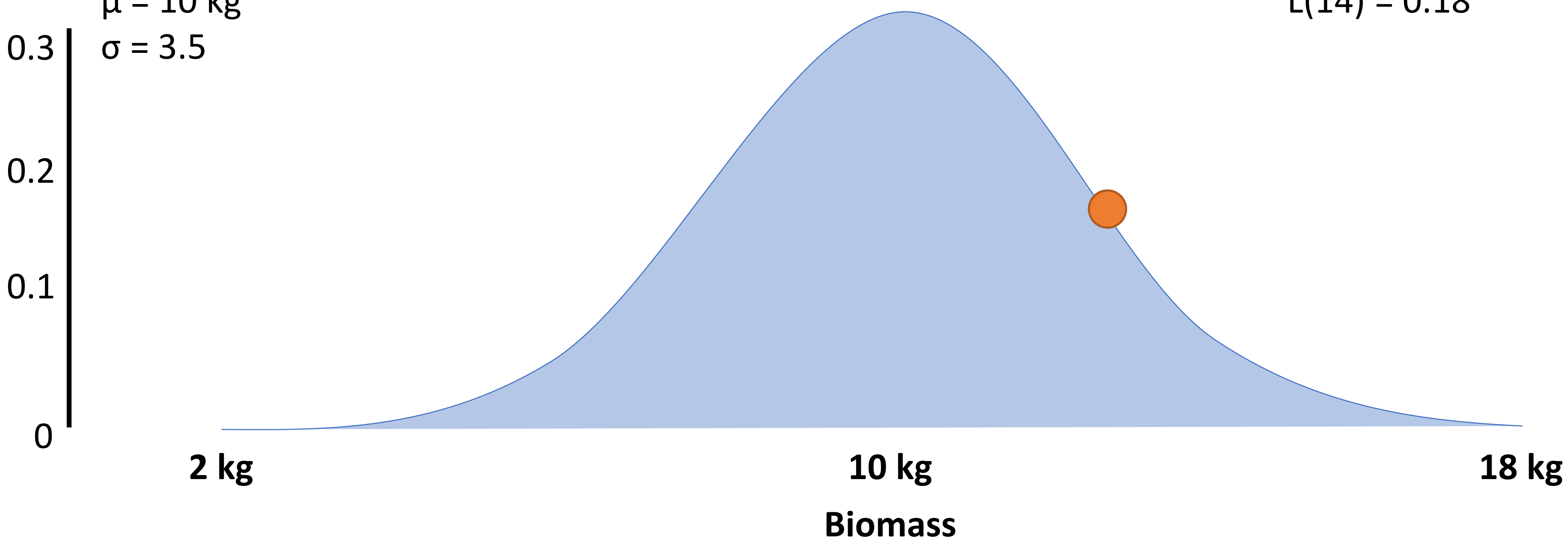


GLM using a normal distribution

Probability vs. **Likelihood**

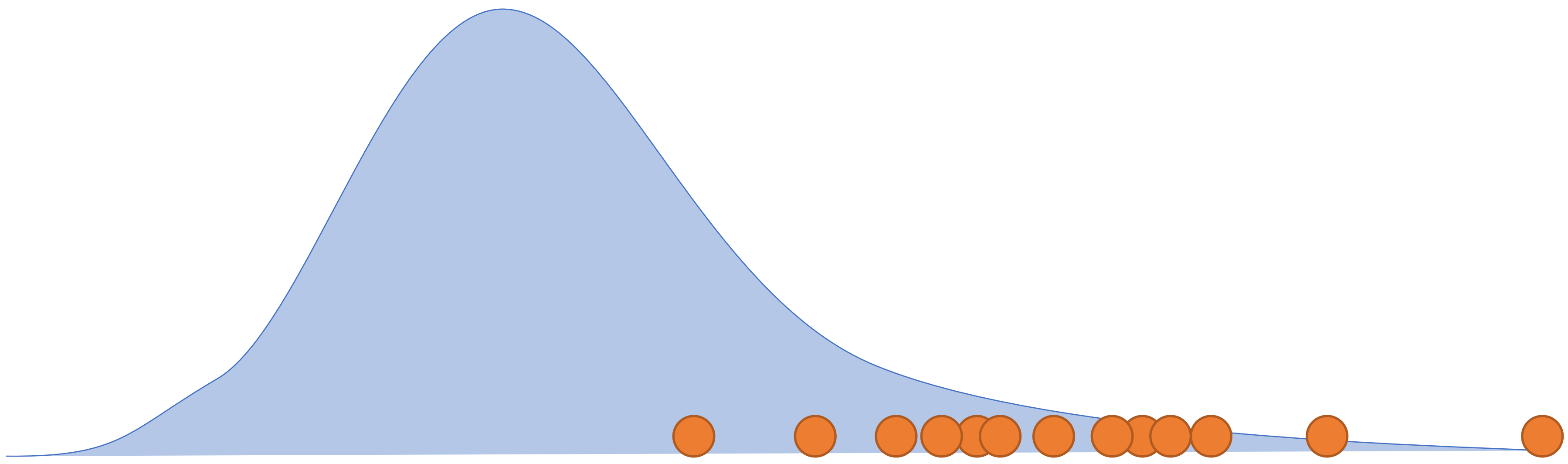
Normal distribution
μ = 10 kg
σ = 3.5

L (14) = ?
L(14) = 0.18



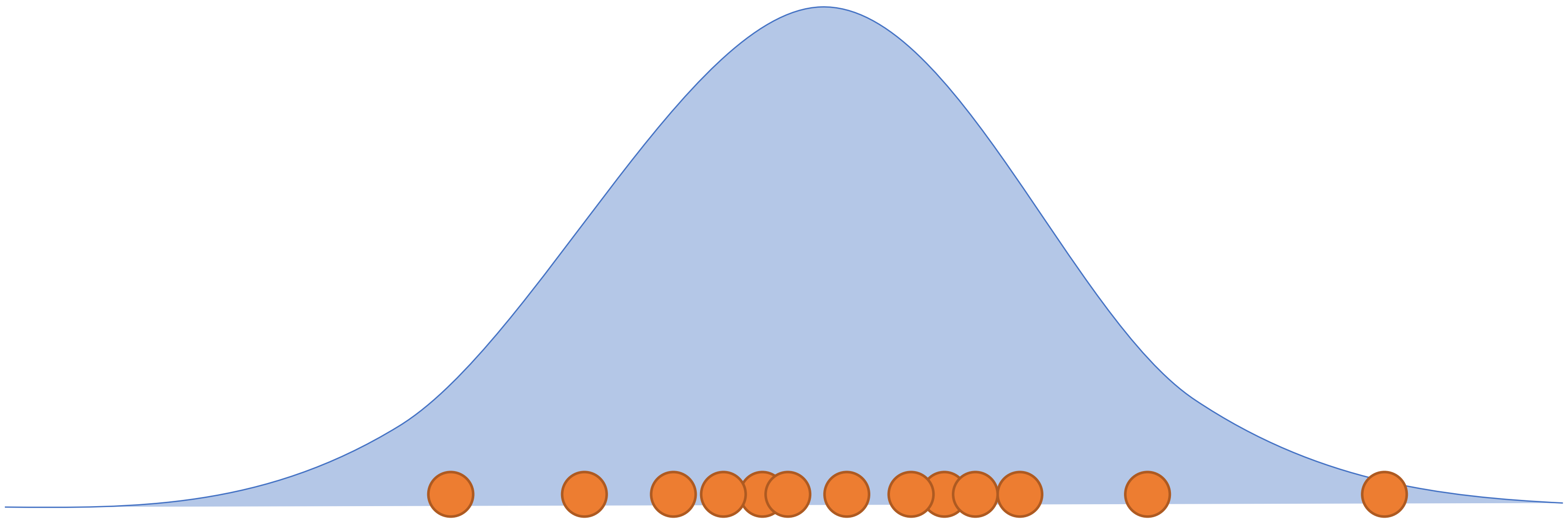
Maximum likelihood for normal distribution

Maximum likelihood = L(1) * L(2) * … L(i)



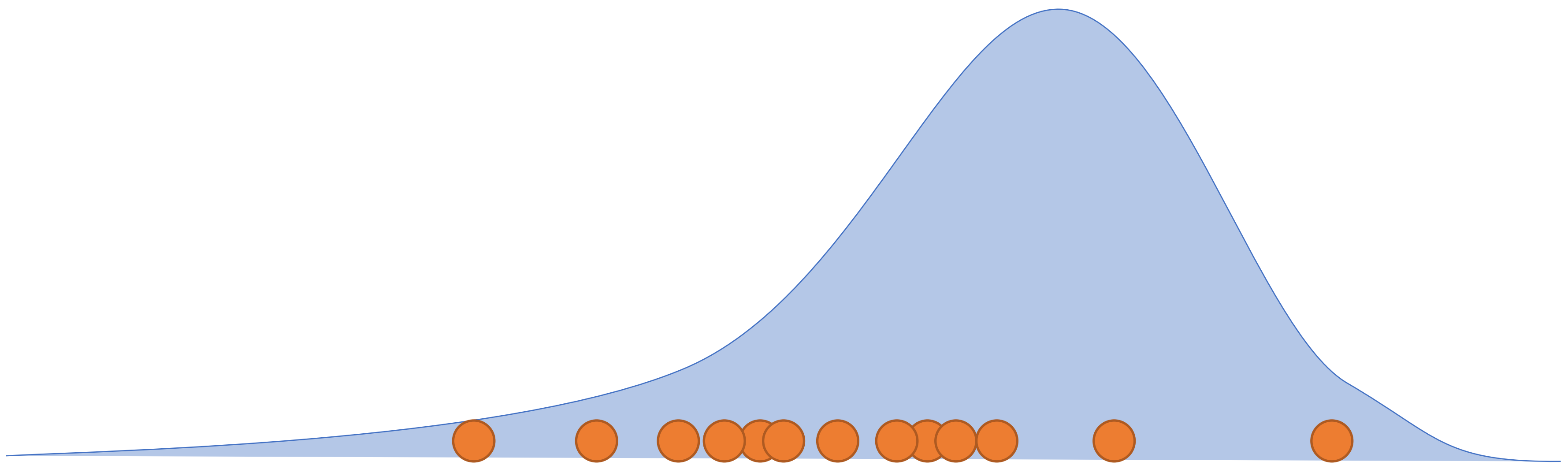
Maximum likelihood for normal distribution

Maximum likelihood = L(1) * L(2) * … L(i)

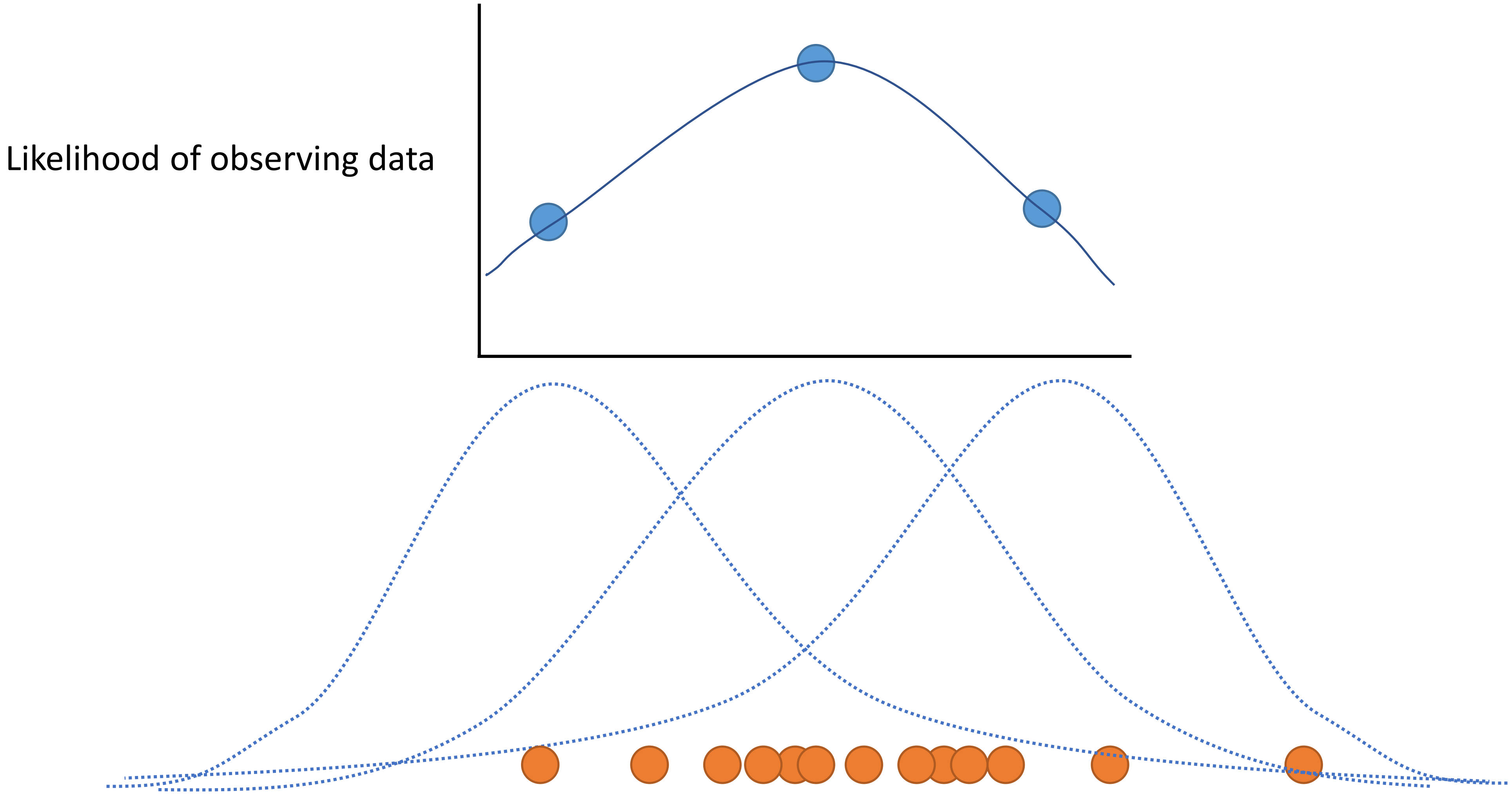


Maximum likelihood for normal distribution

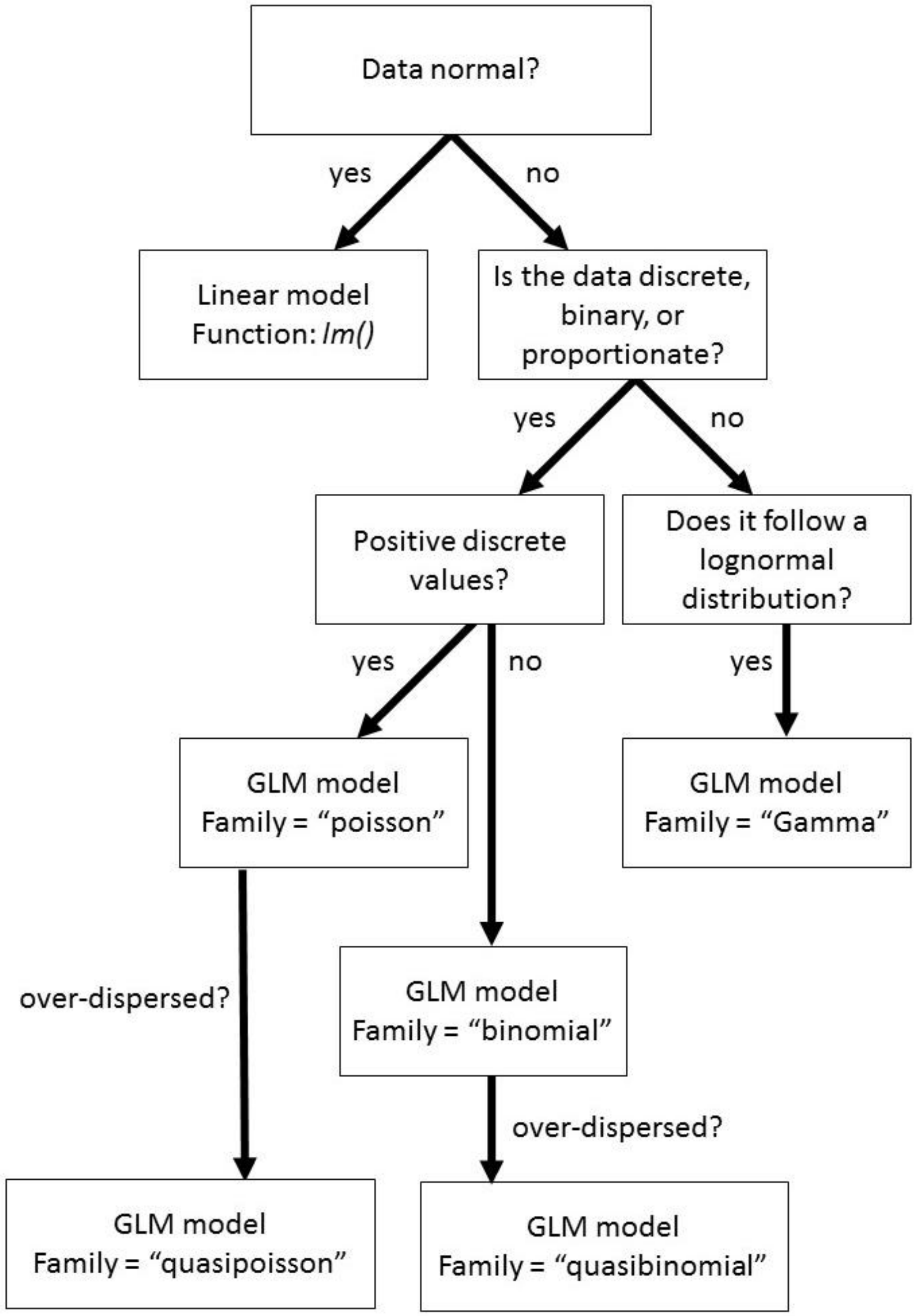
Maximum likelihood = L(1) * L(2) * … L(i)



Maximum likelihood for normal distribution

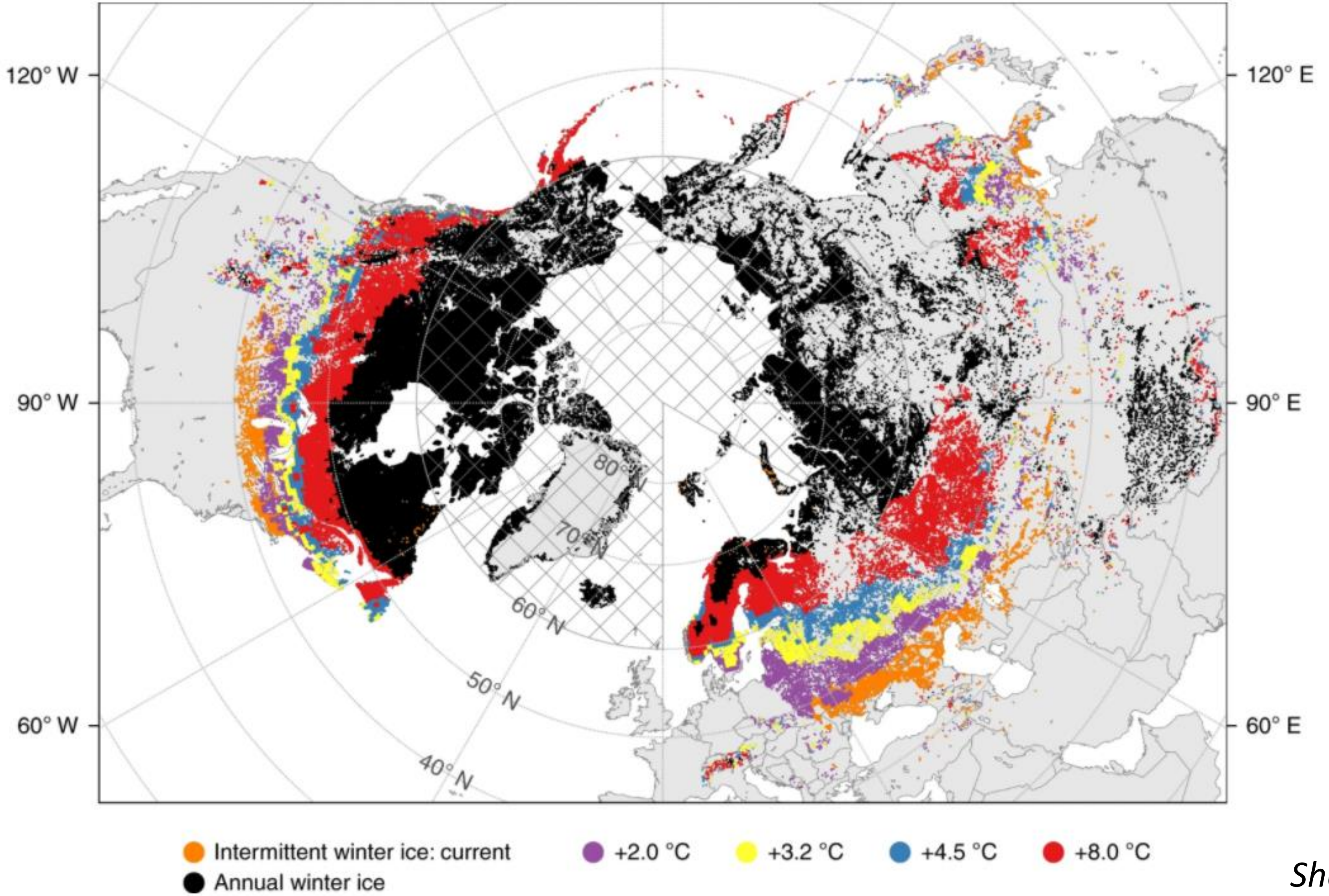


A simple workflow for GLMs



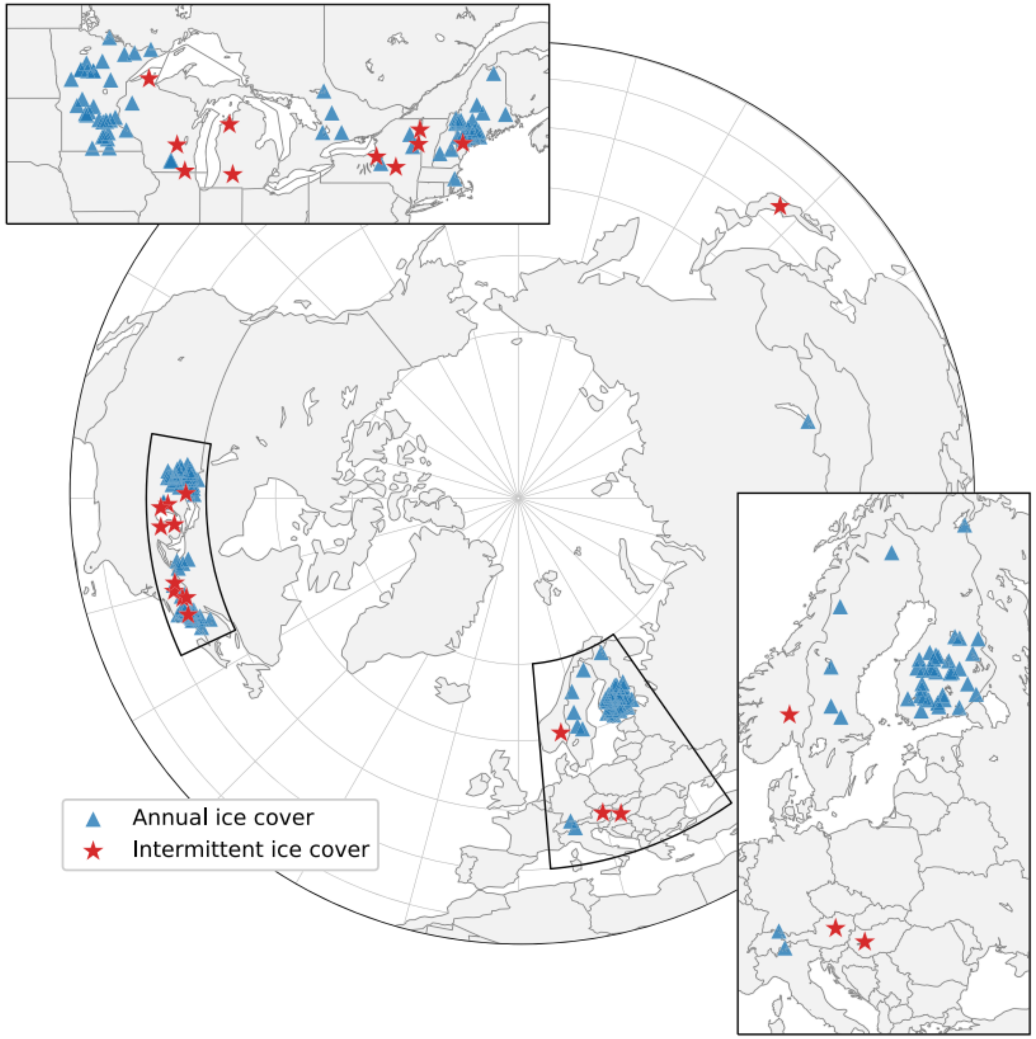
Case Study: Logistic regression to identify extreme events

Lake ice is threatened by climate change



Sharma et al. 2019

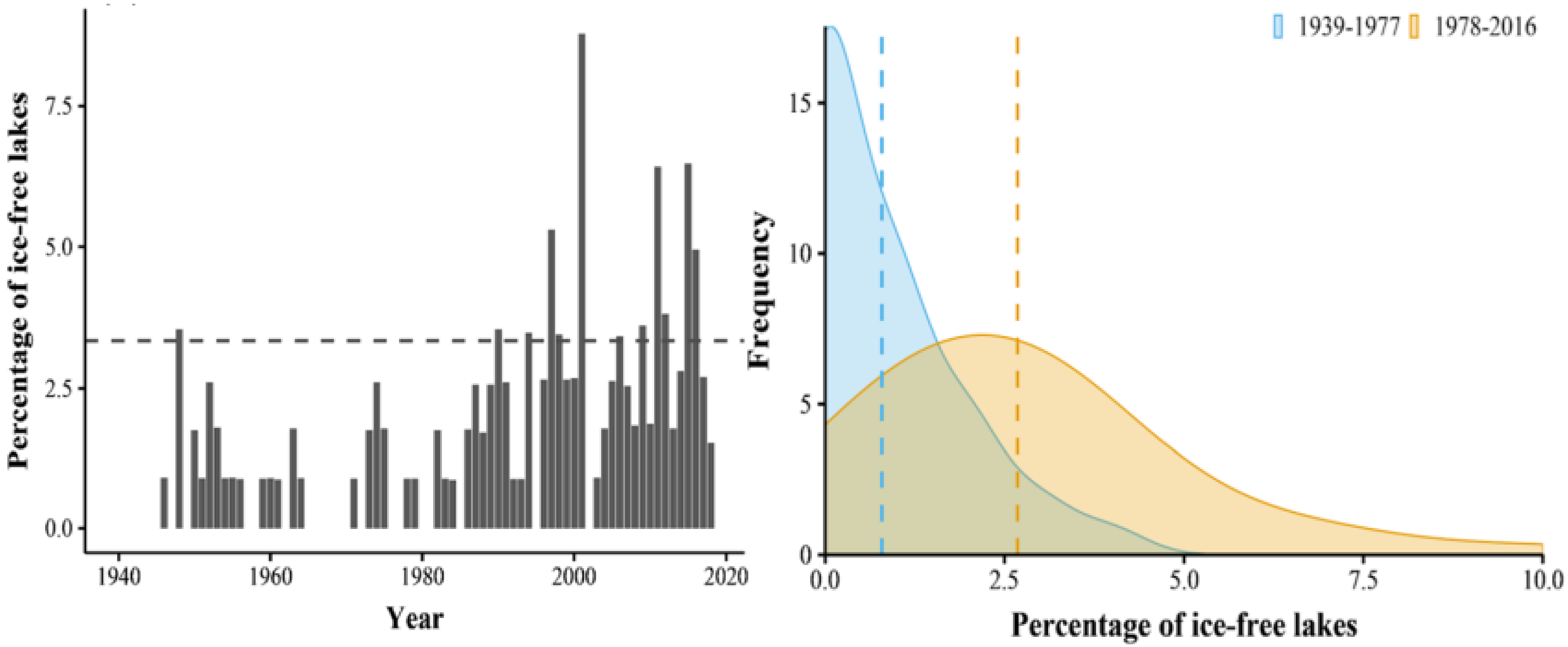
Extreme events on lake ice



- Selected 122 lakes in the Northern Hemisphere
- Tested the frequency of ice-free years over time
- Examined the role of extreme temperatures

Filazzola et al. 2020

A large increase in ice-free coverage

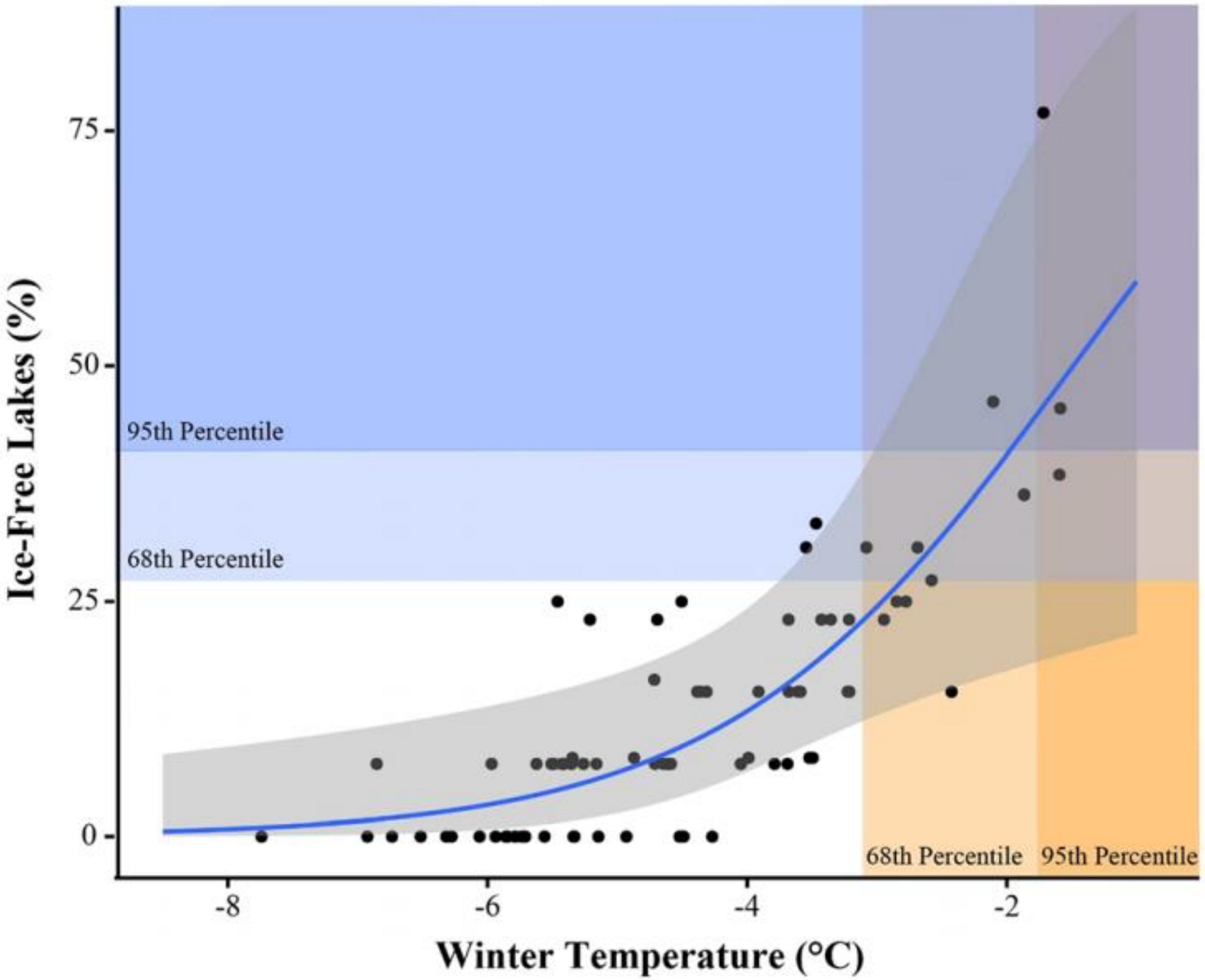


Filazzola et al. 2020

Data of ice-free events

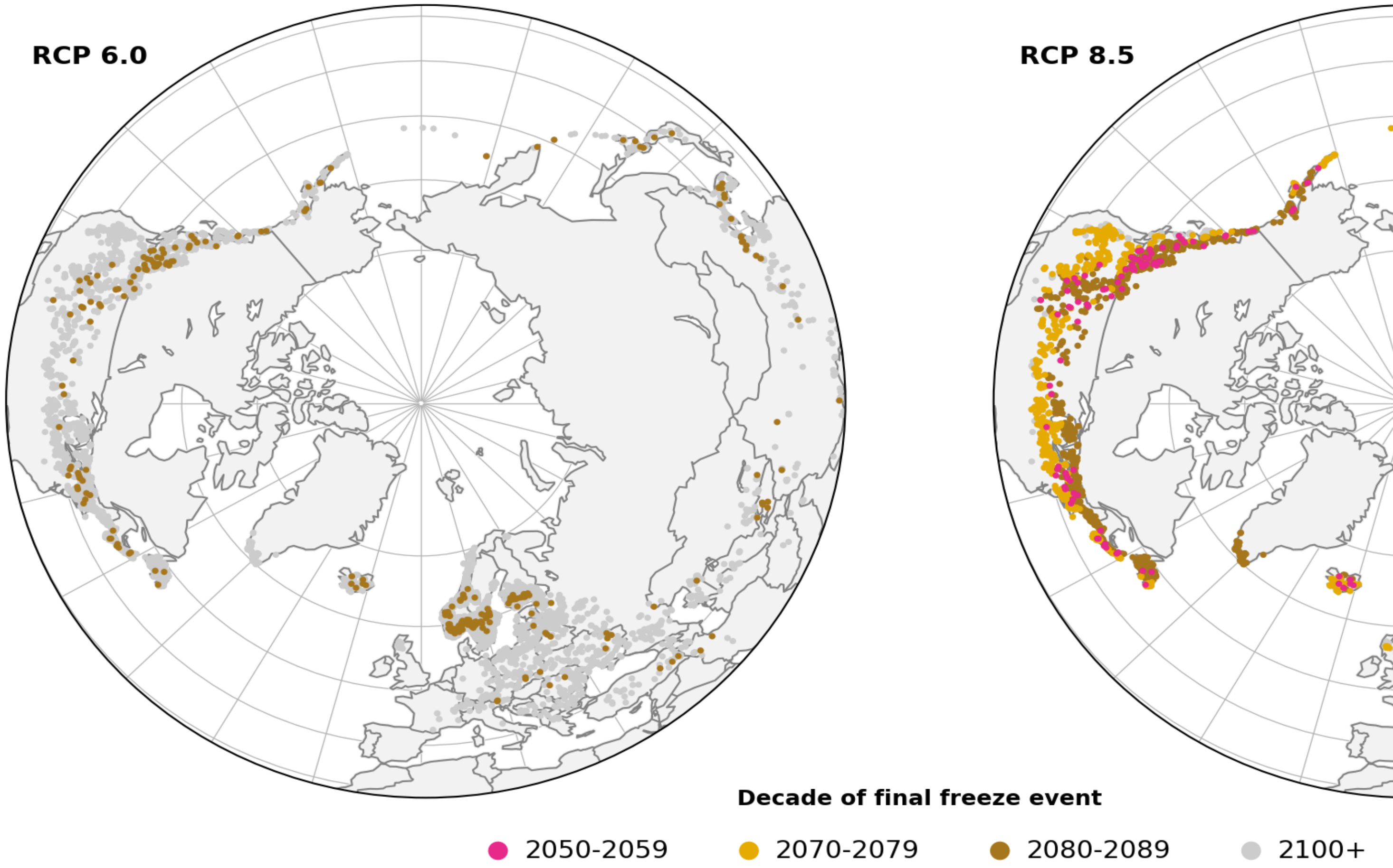
Lake	Year	IceFree	Winter air temperature
Sebago Lake	1995	1	-1
Sebago Lake	1996	0	-2.5
Sebago Lake	1997	0	-2.9
Sebago Lake	1998	0	-3.1
George Lake	1995	0	-2.1
George Lake	1996	1	-1.6
George Lake	1997	1	-0.5

Extremes in temperature cause extremes in ice coverage



Filazzola et al. 2020

The future loss of ice coverage



Sharma et al. in review

UNIVERSITY OF TORONTO

Thank you!

https://www.filazzola.info/

https://github.com/afilazzola/CUELogisticRegression