# Generalized Linear Models (GLMs)

Dr. Alessandro Filazzola & Dr. Christopher Lortie

BIOL 5018

# Outline & Learning Outcomes

**Part A – Conducting GLMs**

- Describe the uses of GLMs relative to linear regression
- Be able to conduct a GLM in R
- Gain a basic understanding of the mechanics and outputs of a GLM

**Part B – Prediction**

- Describe how GLMs can be used for predictions with new data
- Explain the approach for optimizing predictions

**Part C – Post-hoc analyses and mixed models**

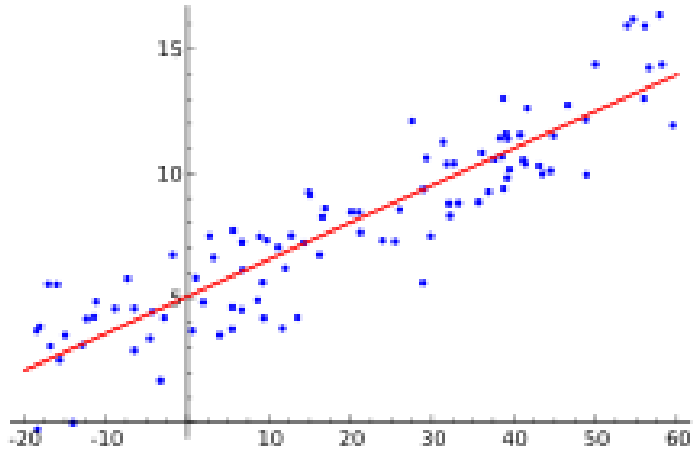
- Understand the differences between GLMs and GLMMs
- Explain the purpose of post-hoc analyses and when they may be used
- Be able to conduct a GLMM and post-hoc analysis in R

Part A

# 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

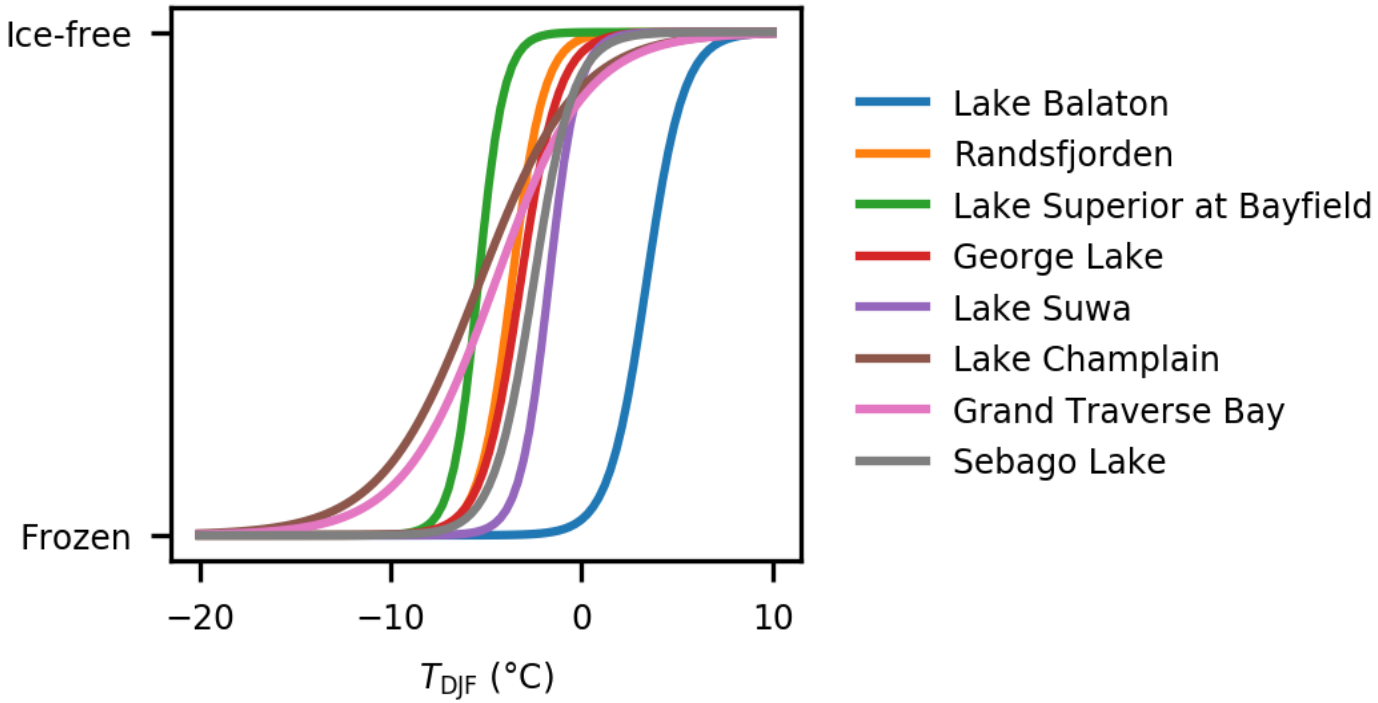
2) The variance of the response variable depends on the mean

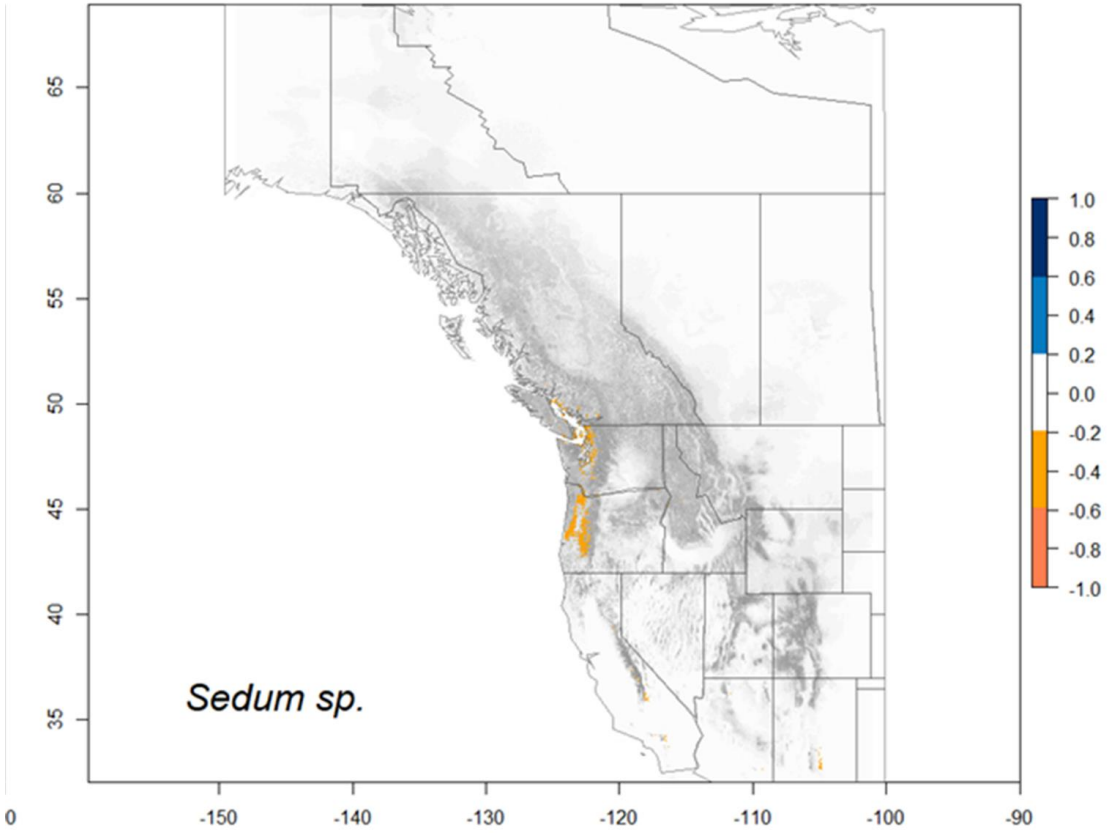
3) GLMs can fit to particular distributions without transformations. The benefits of this include i) the homogeneity of variance does NOT need to be satisfied and ii) the response variable does not need to be changed.

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

# Examples of GLMs

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

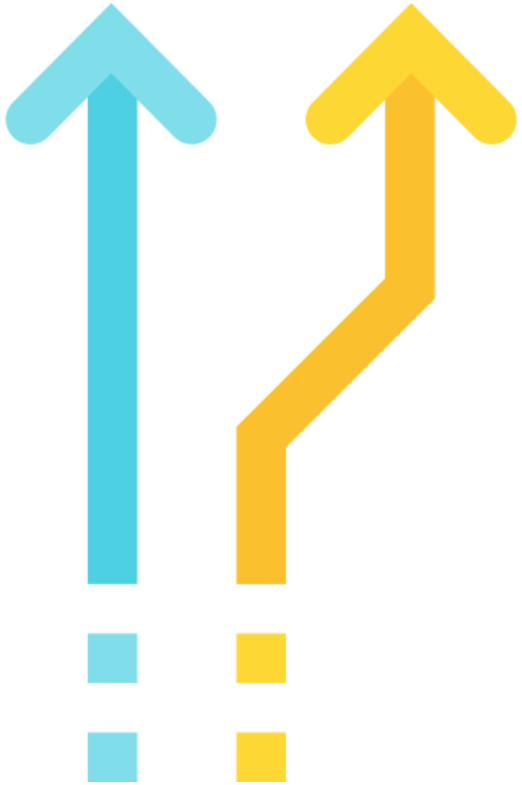


# Parallels with linear regression

Both logistic and linear regression can:

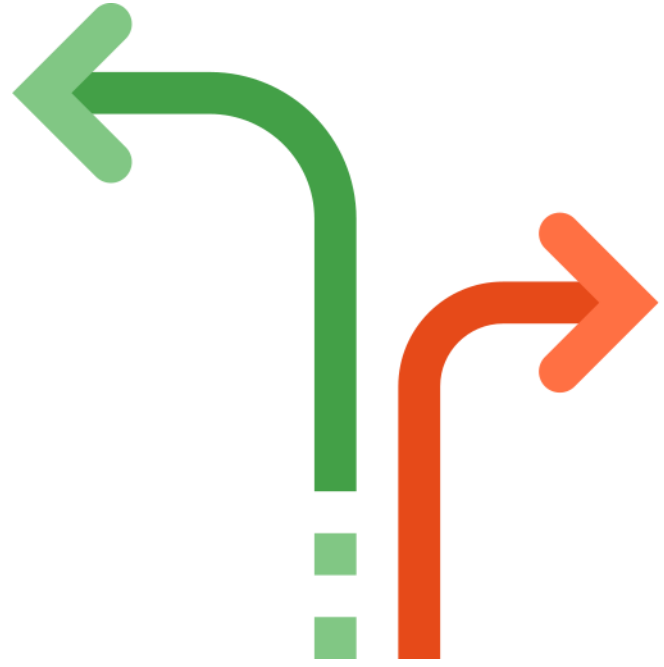
- Test for significance of a predictor
- Calculate an effect size
- Use continuous or categorical predictors
- Can predict values for new values



# Differences with linear regression

Logistic regression differs from linear regression in that:

- Response variable can be complex (e.g., binary, discrete, continuous)
- There is no “true” R2 or residuals
- Uses Maximum Likelihood instead of Sum of Squares

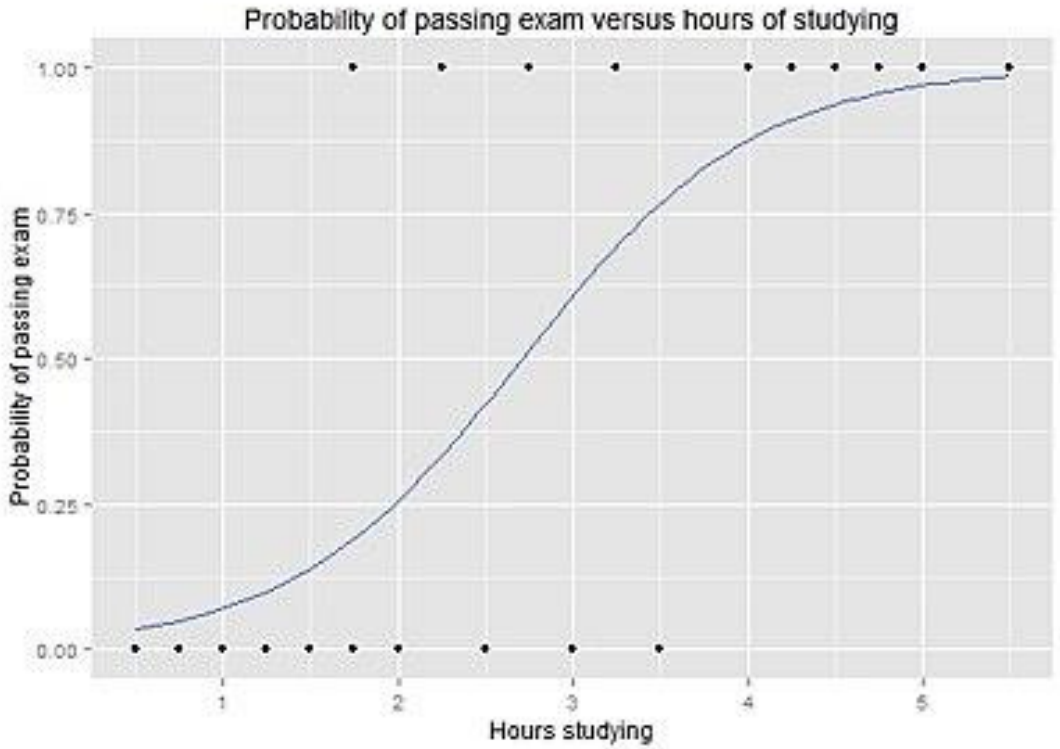


# Understanding GLMs

1) Fitting a line

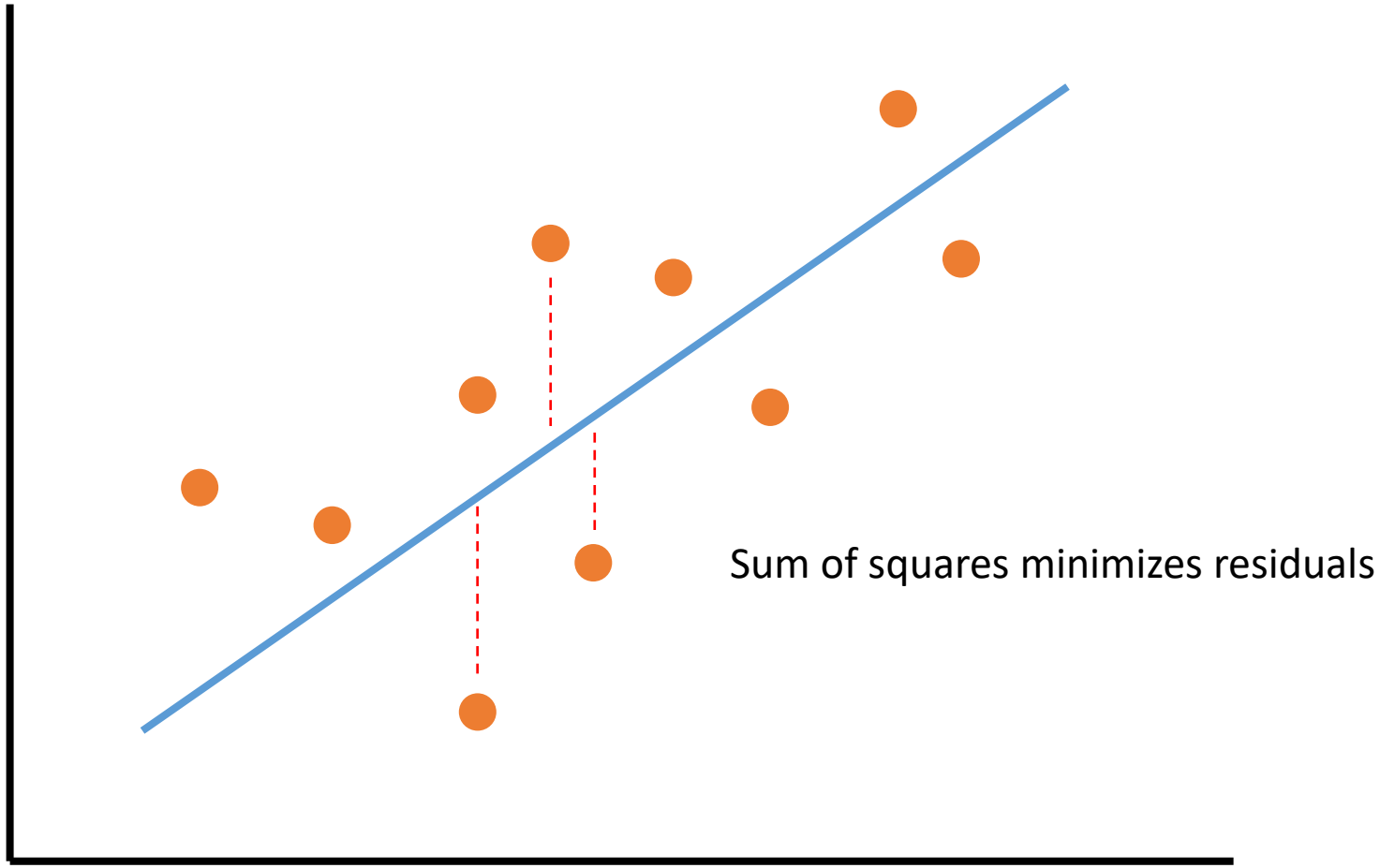
2) The outputs

3) Calculating fit

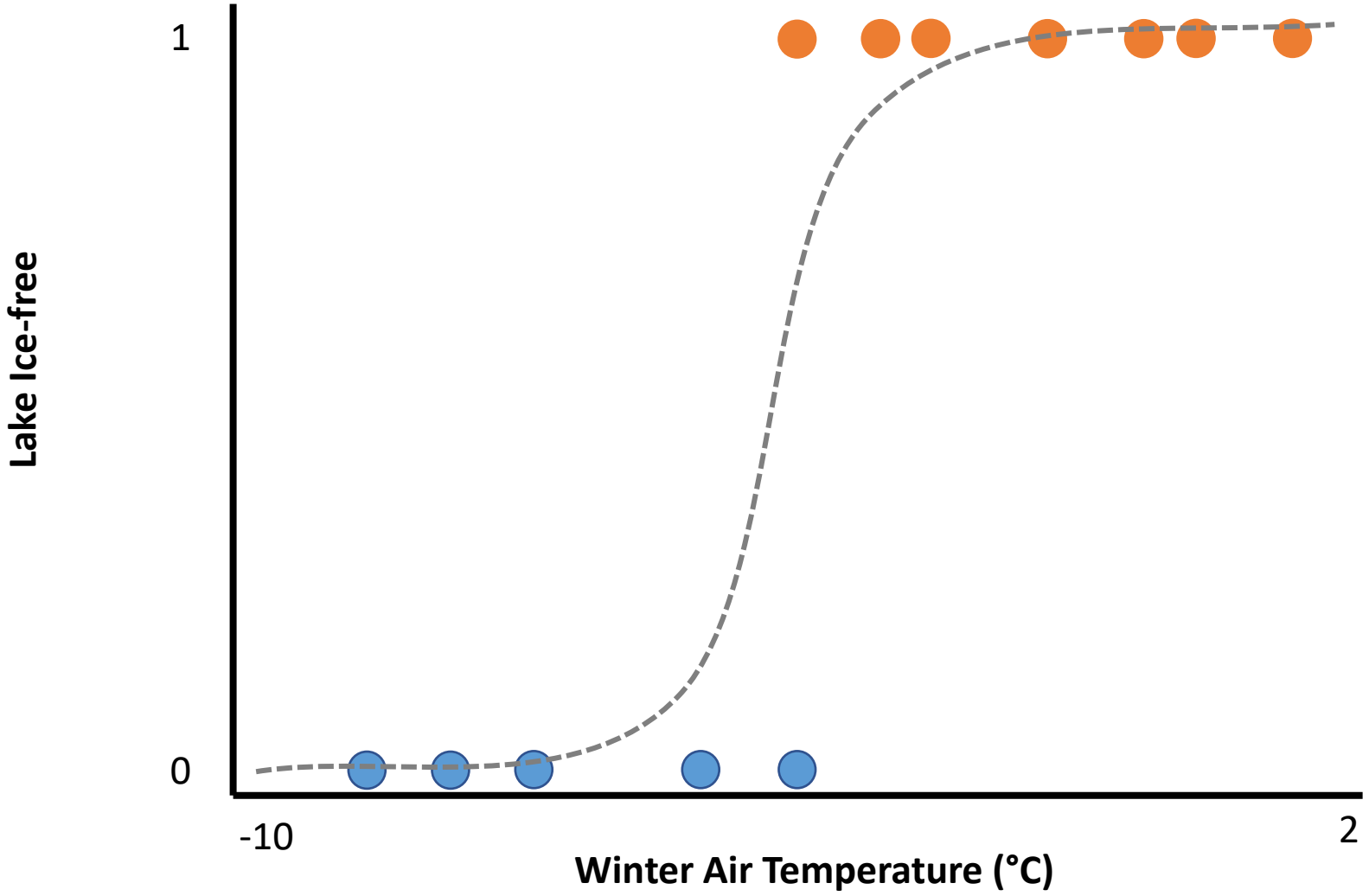



*Wikimedia.com*

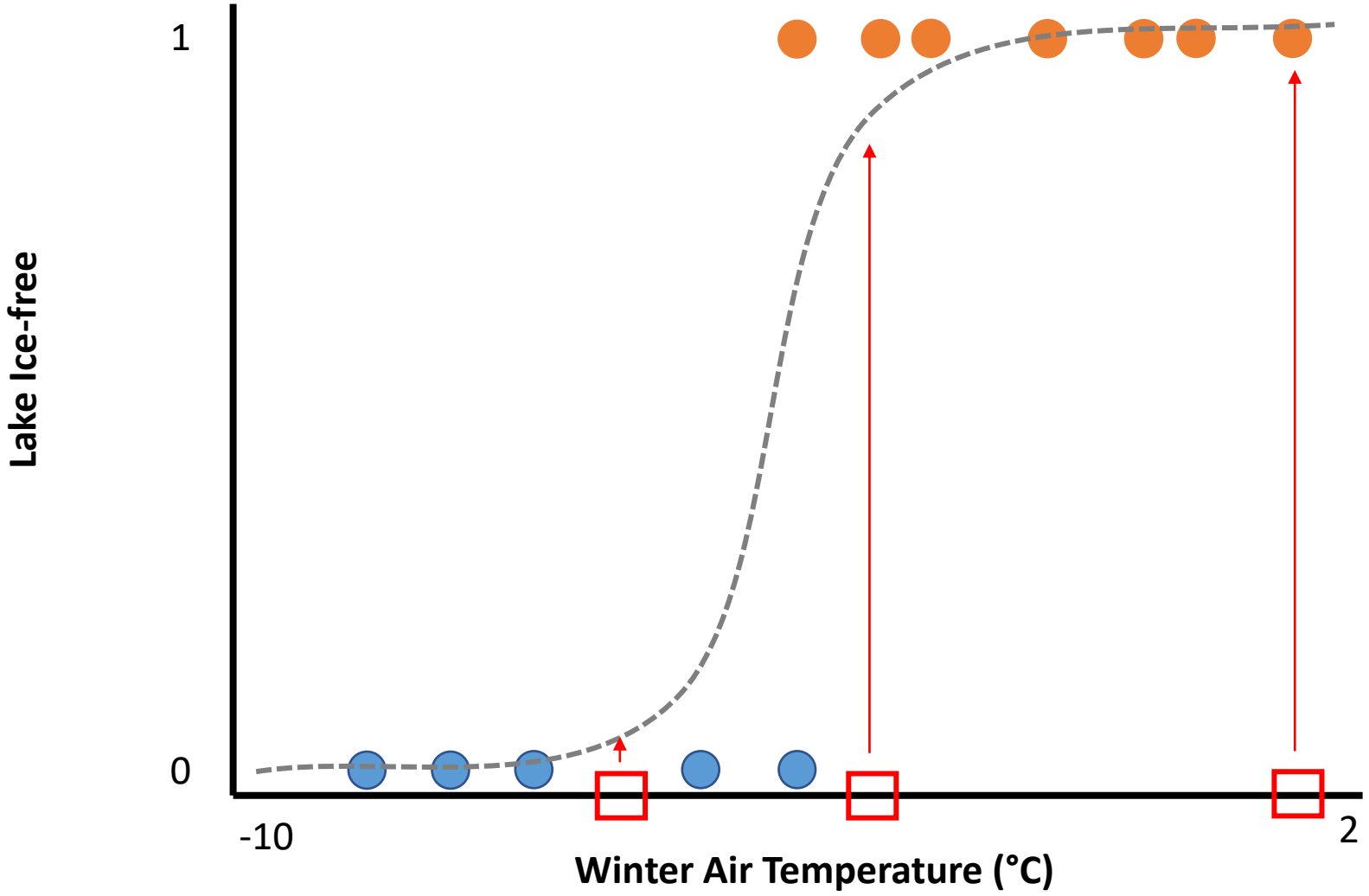
# 1) Fitting a line: linear regression



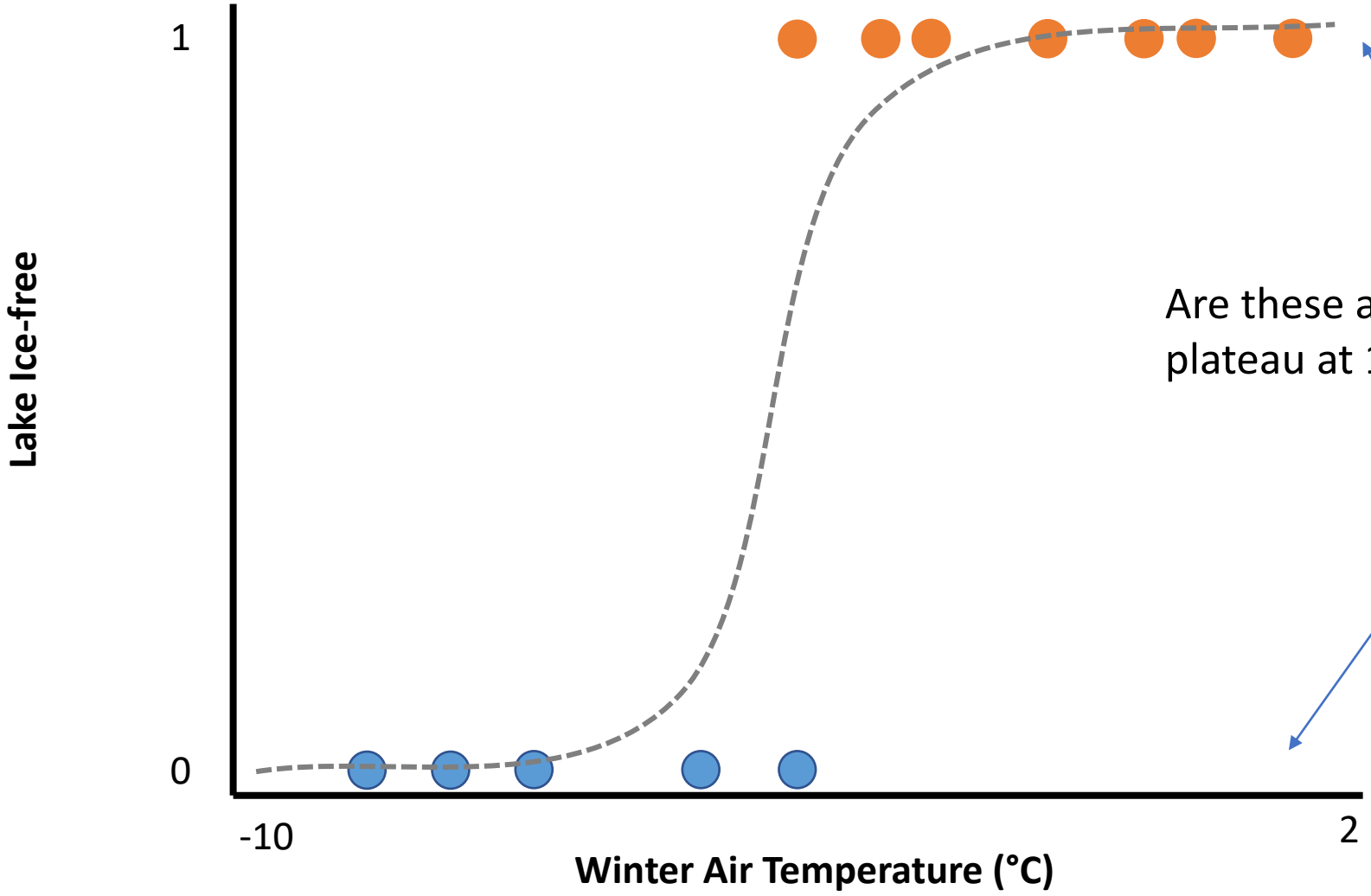
# 1) Fitting a line: logistic regression



# 1) Fitting a line: logistic regression



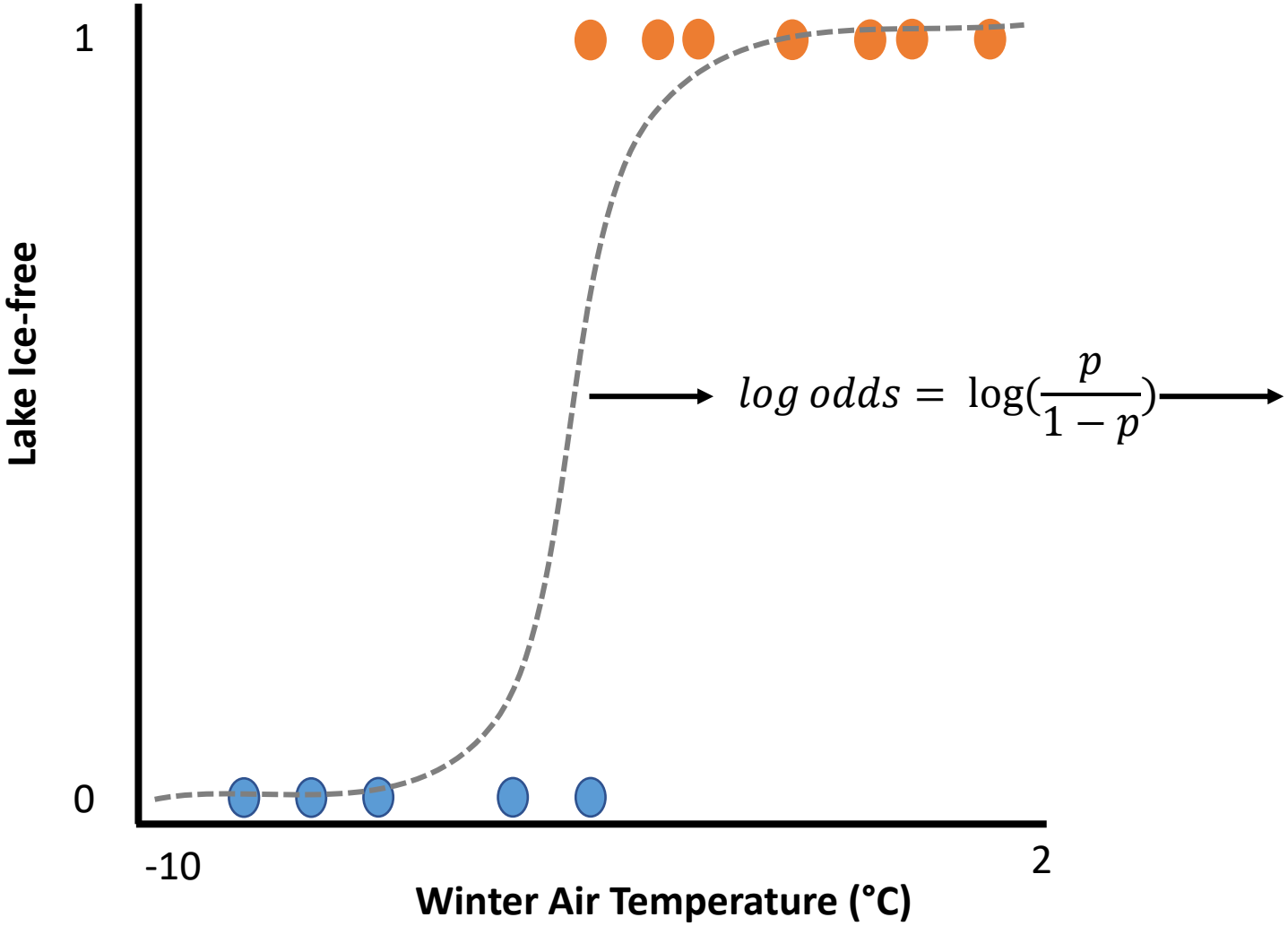
# Question

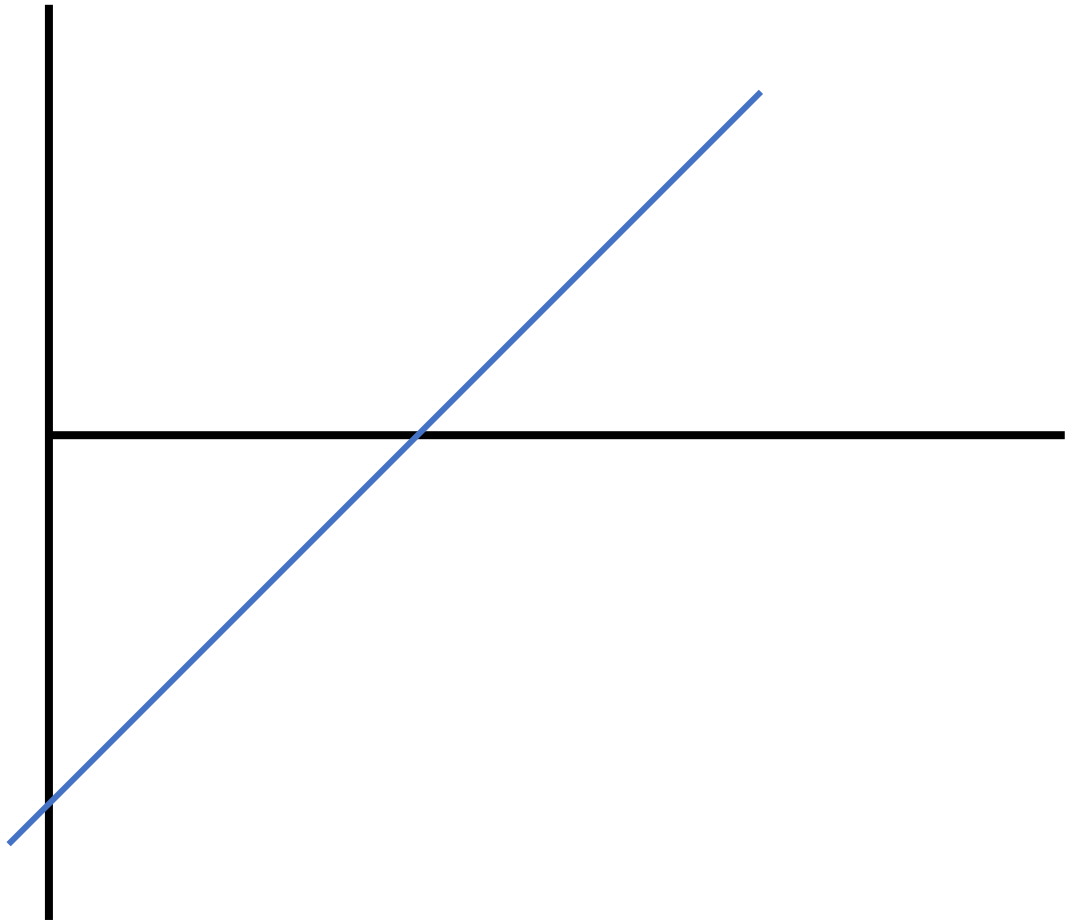



Are these asymptotes or to they plateau at 1 and zero respectively?

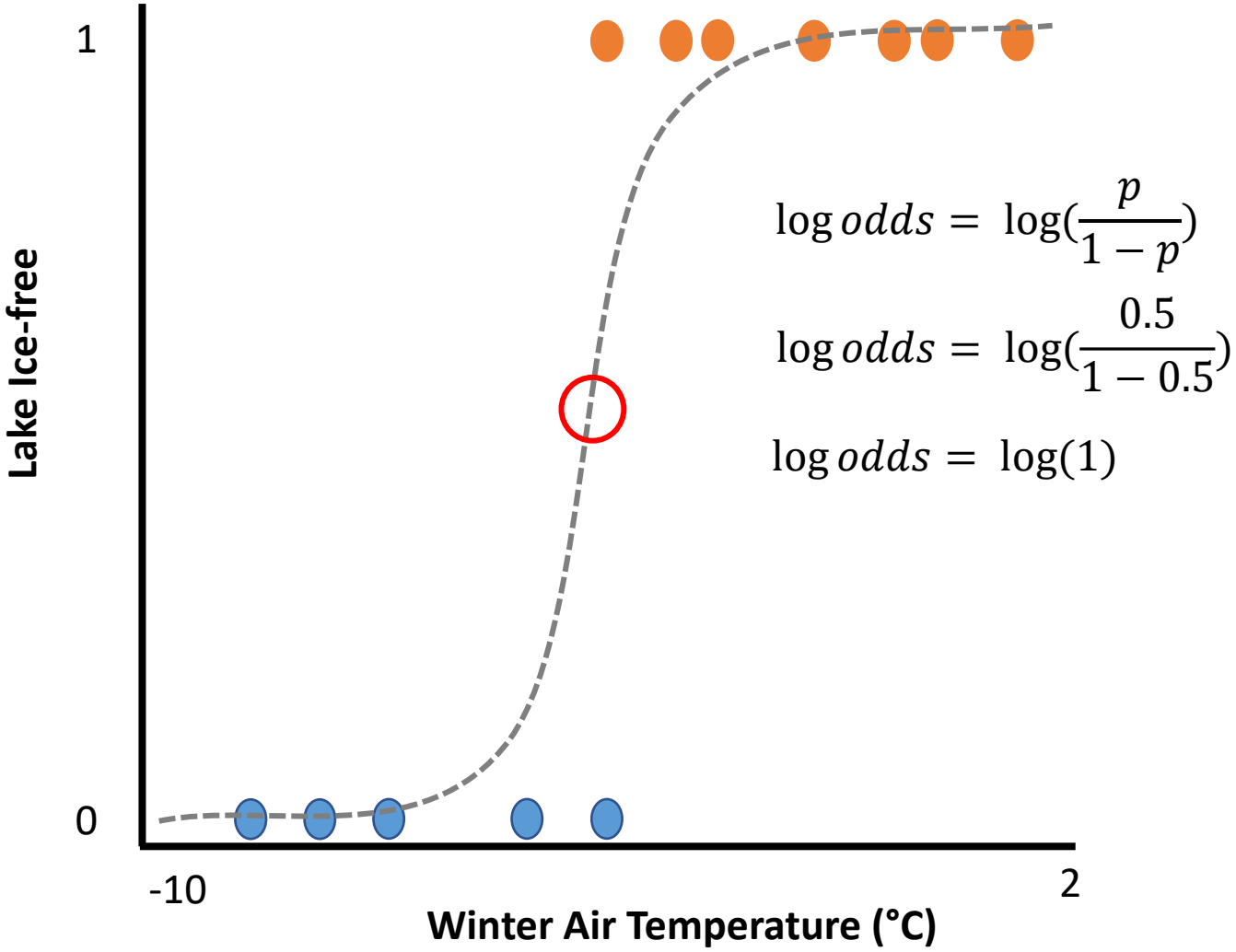
$$log\ odds = \log(\frac{p}{1-p})$$

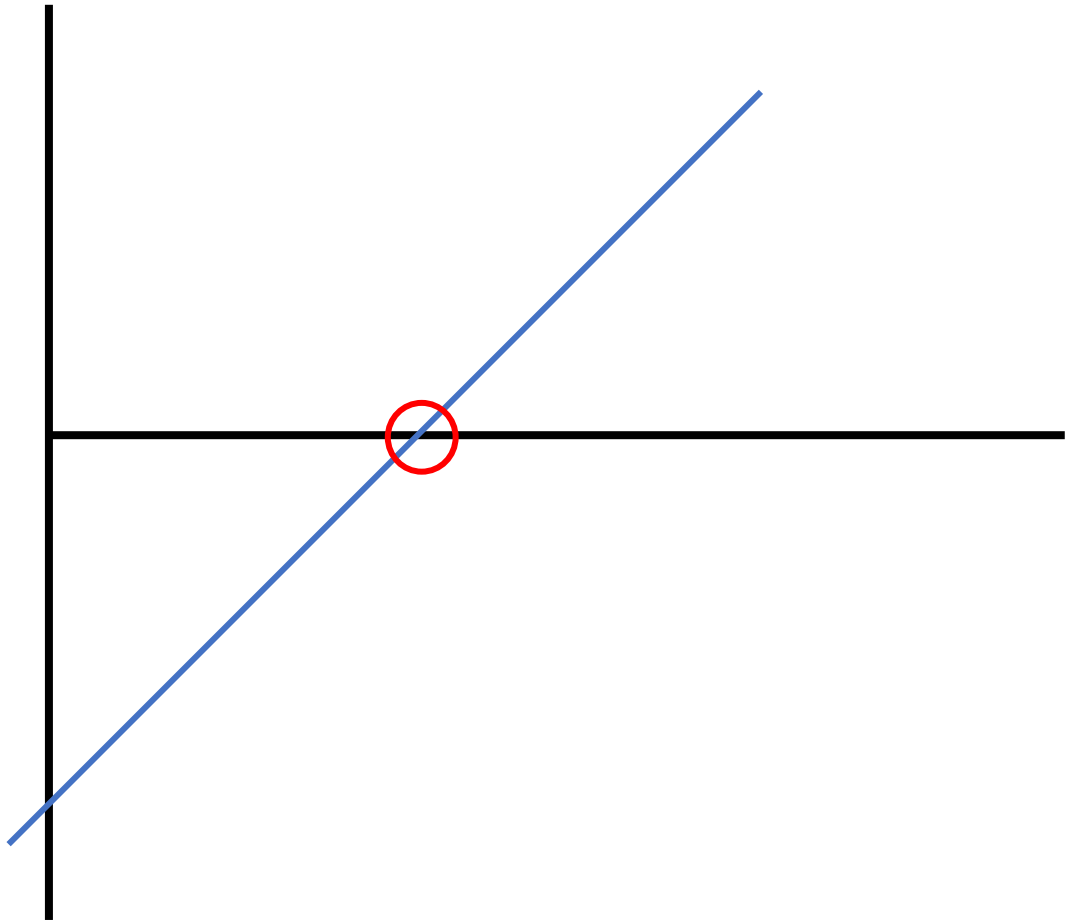
# 1) Putting the *linear* in generalized linear model



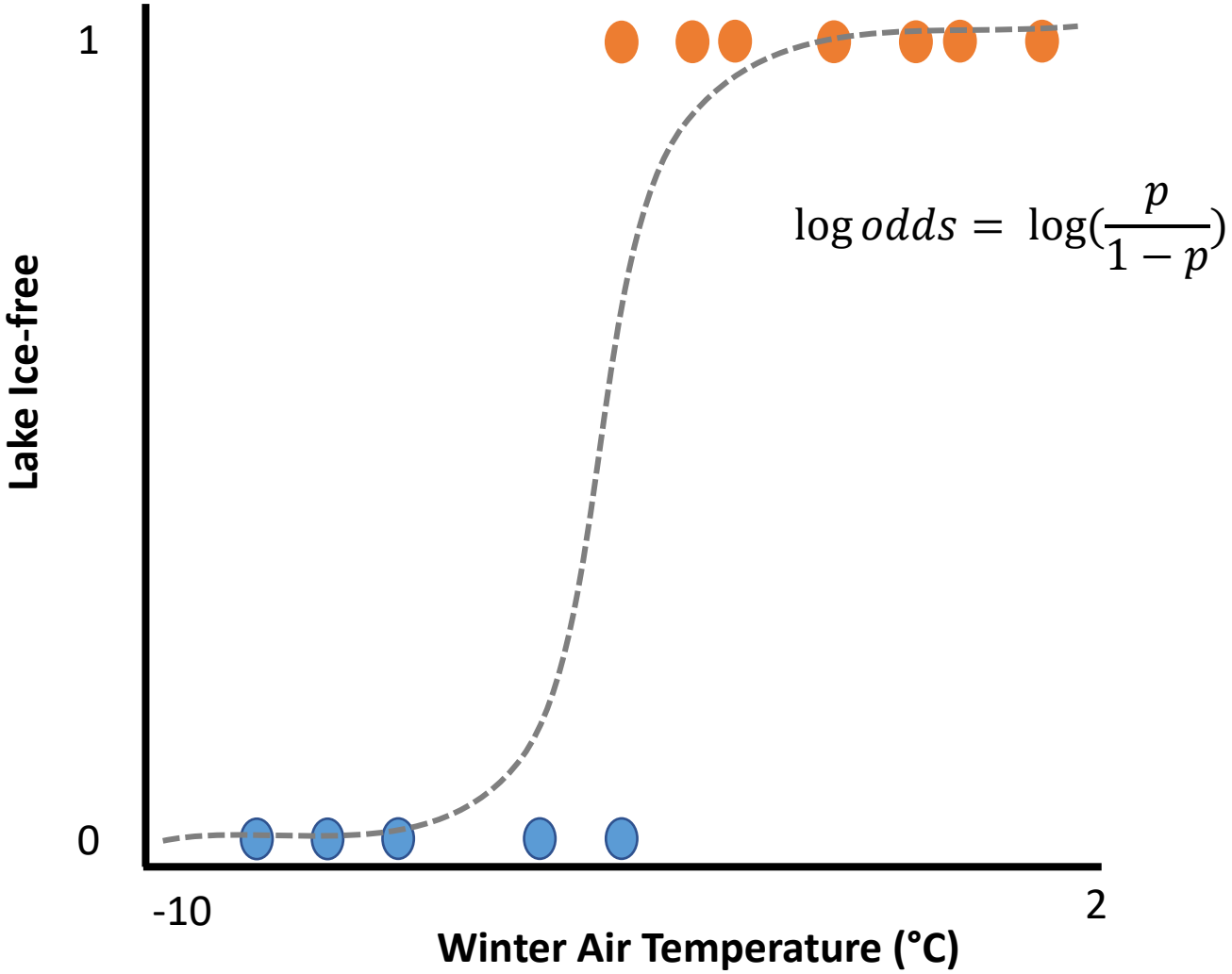
# 1) Putting the *linear* in generalized linear model

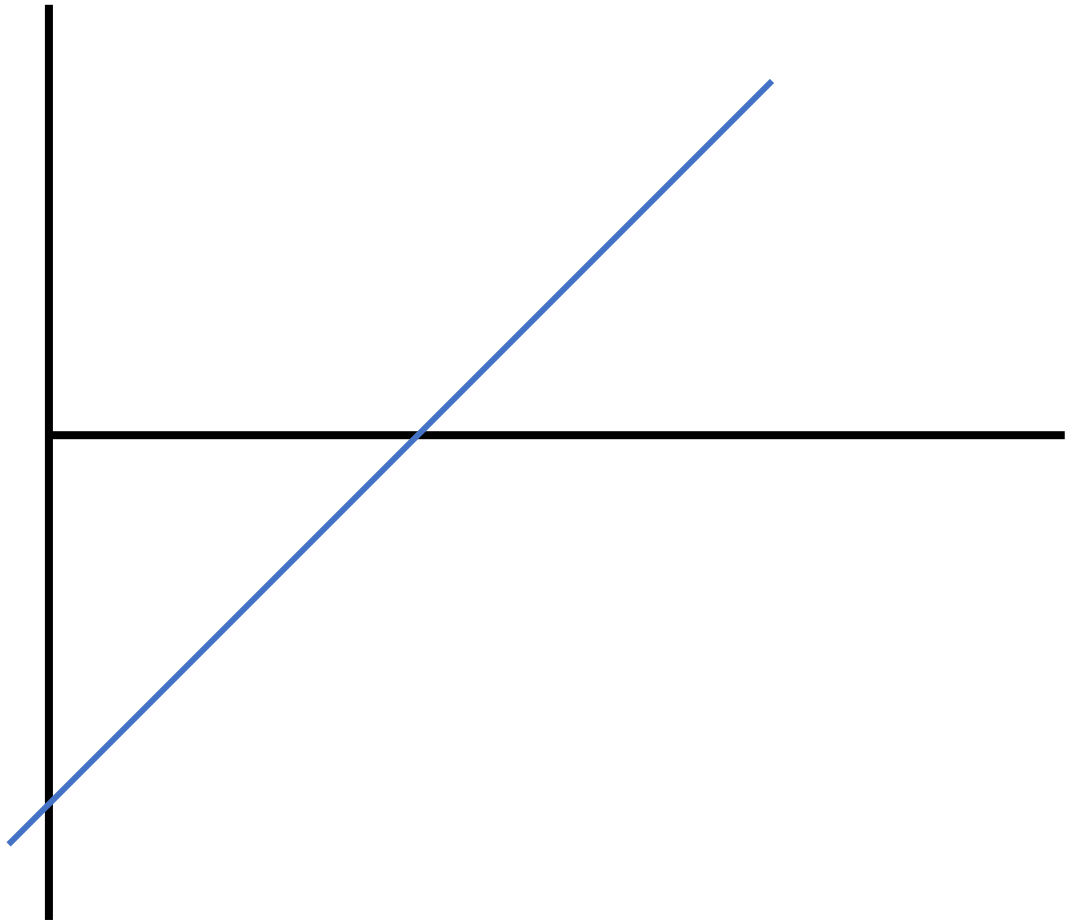


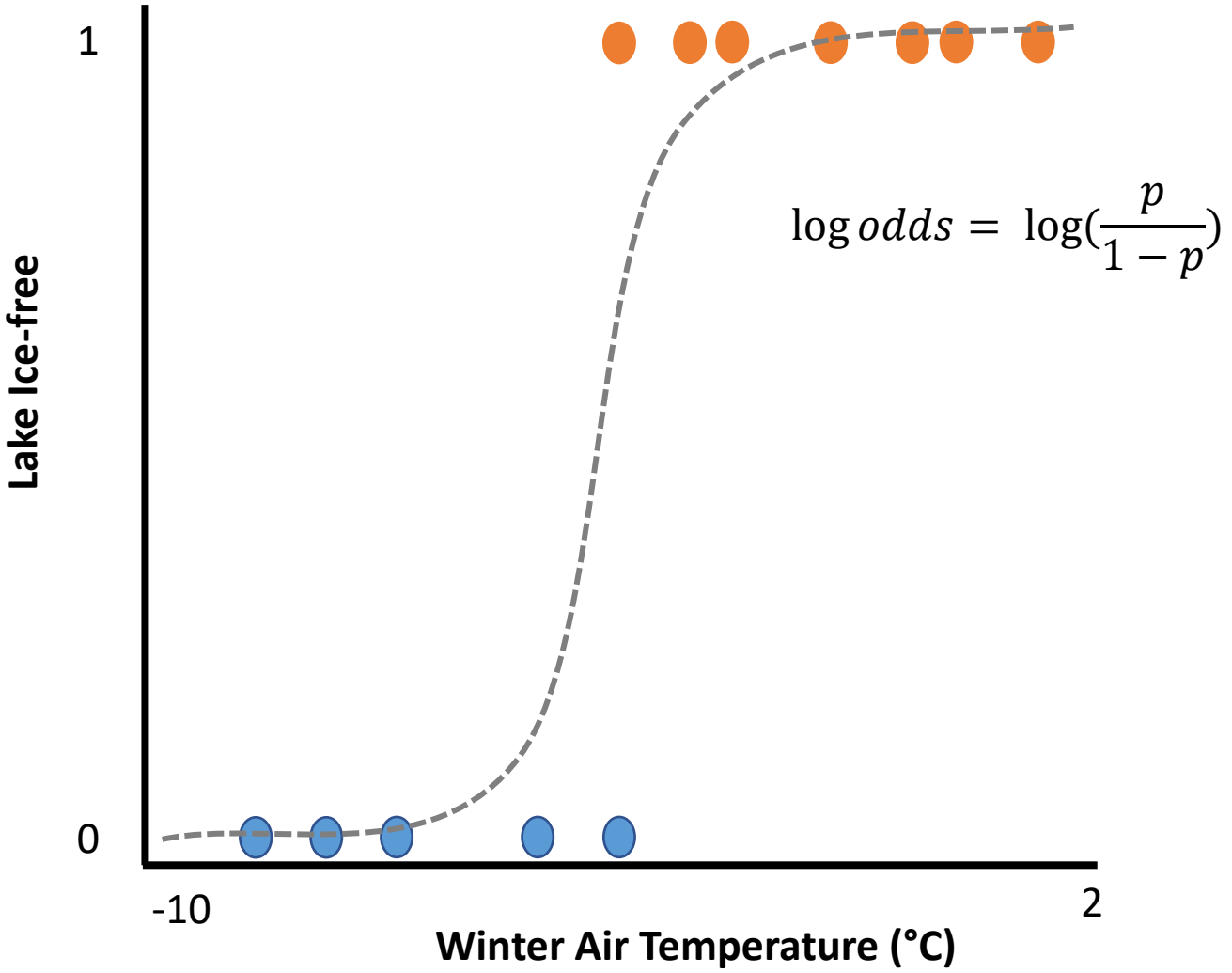
**This is the slope co-efficient from the R Output**

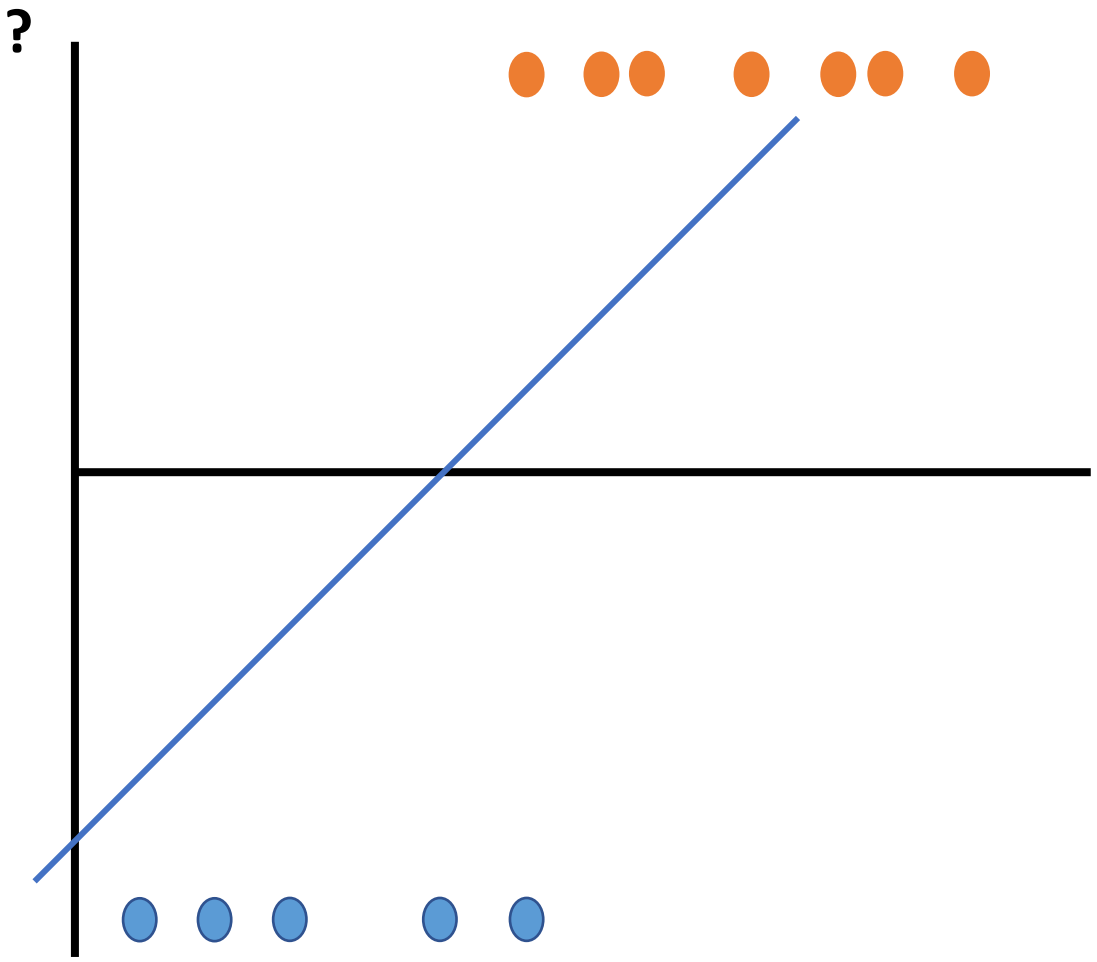
# 1) Putting the *linear* in generalized linear model



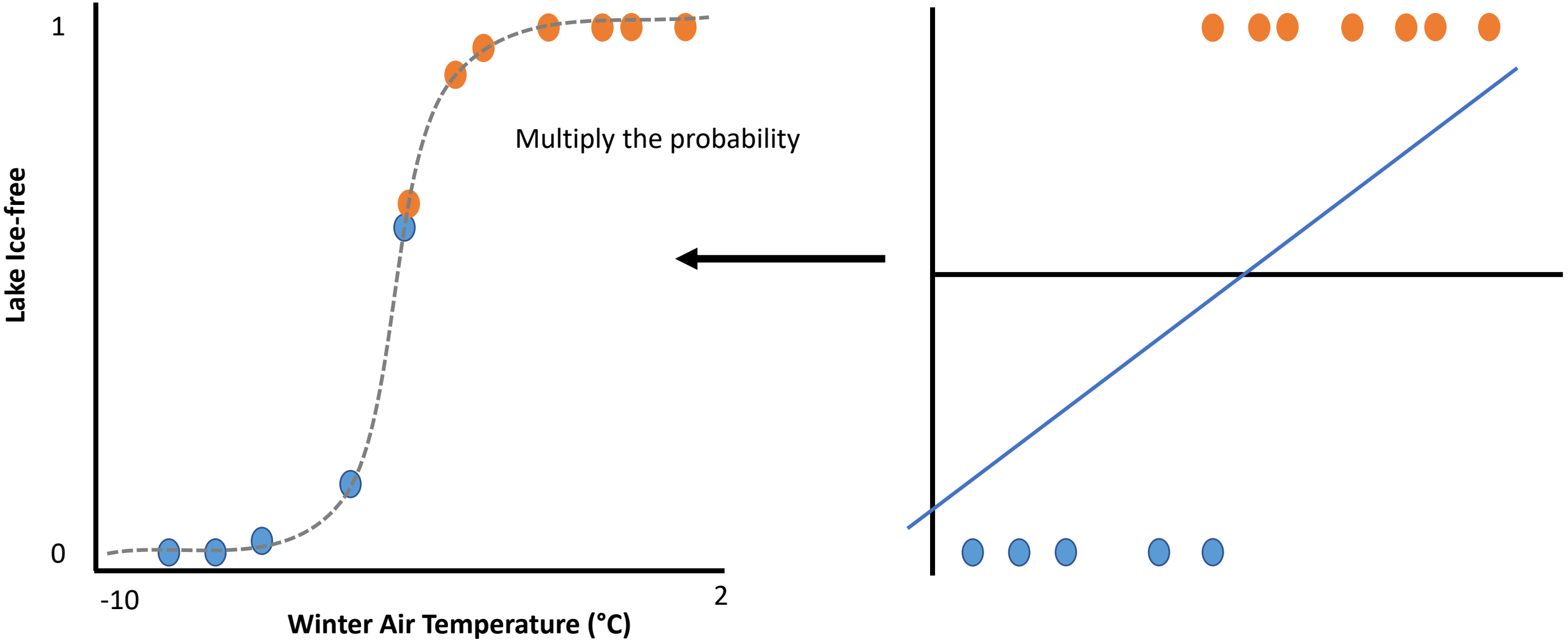



# 1) Putting the *linear* in generalized linear model

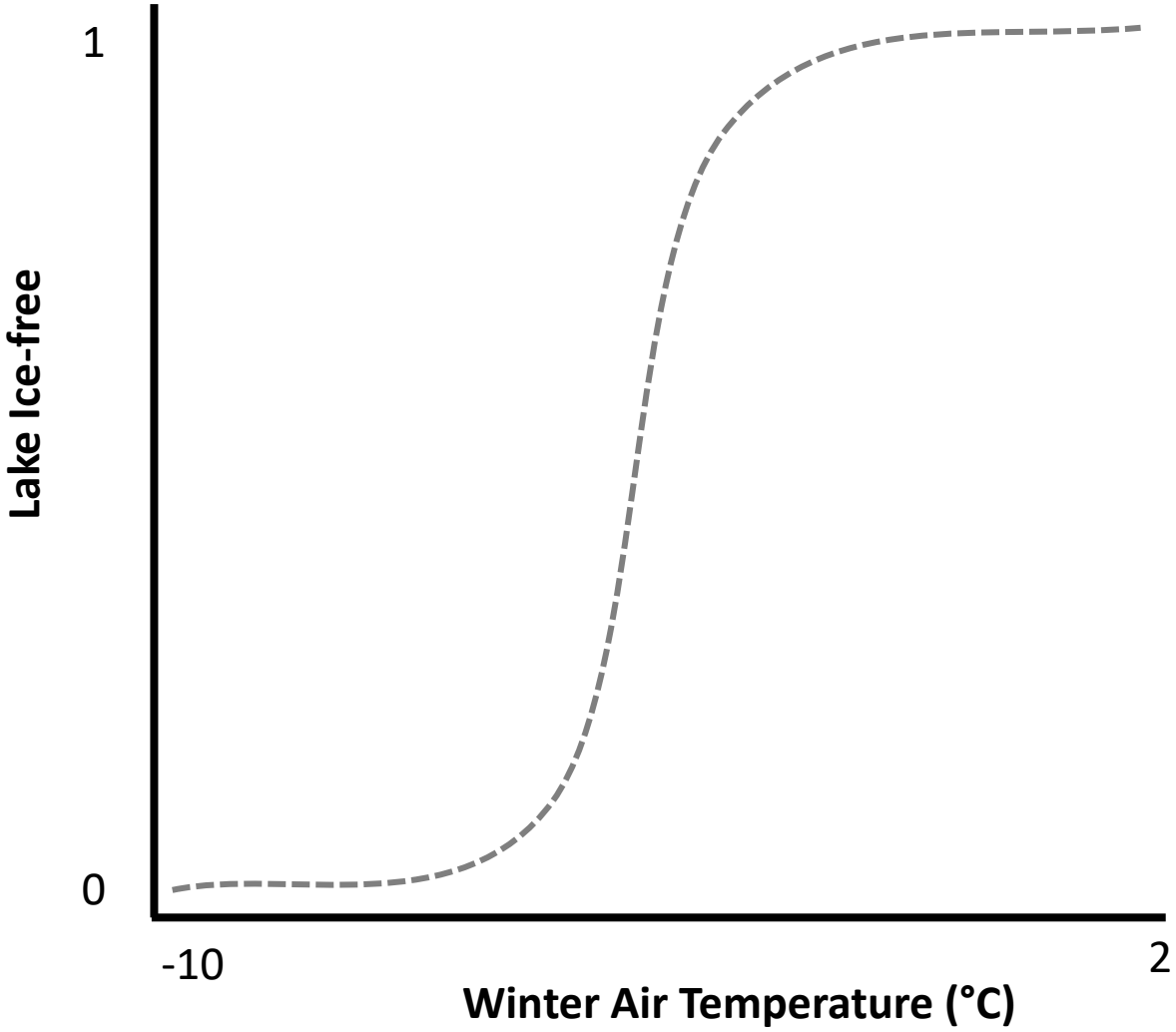


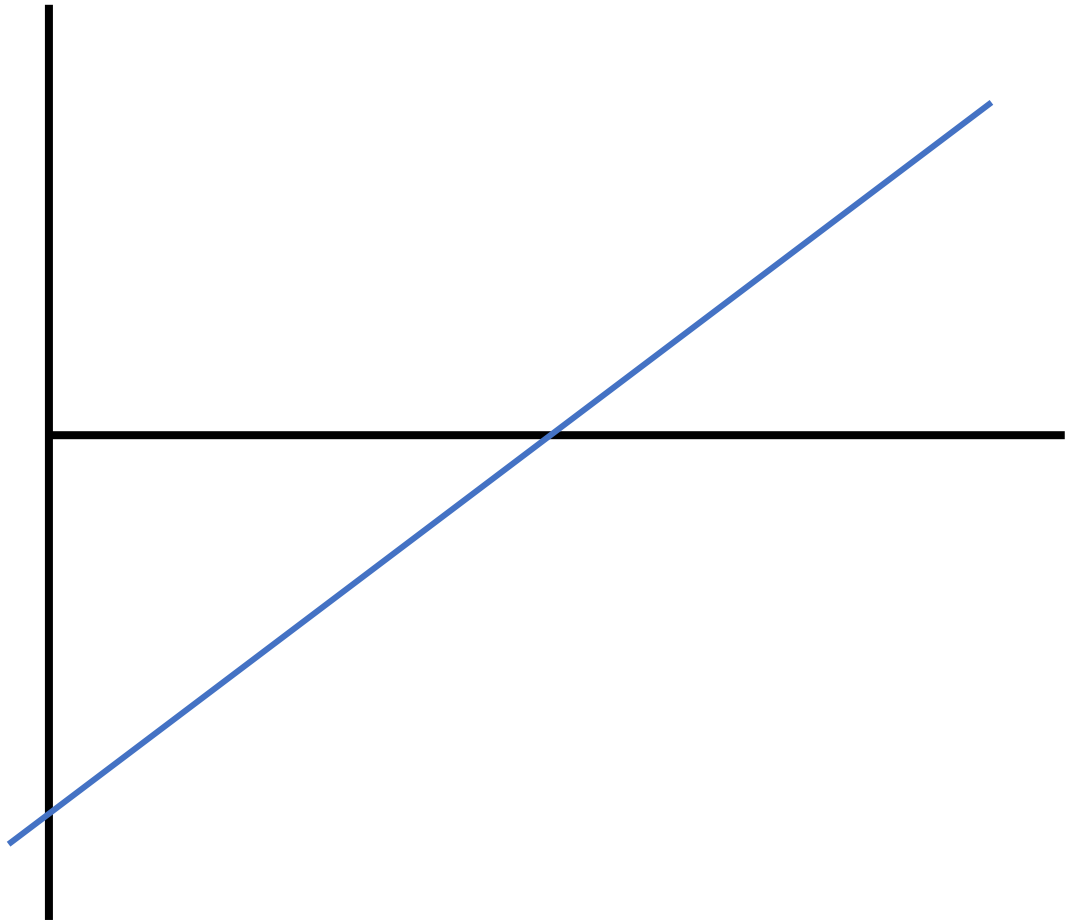



# 1) Fitting the model

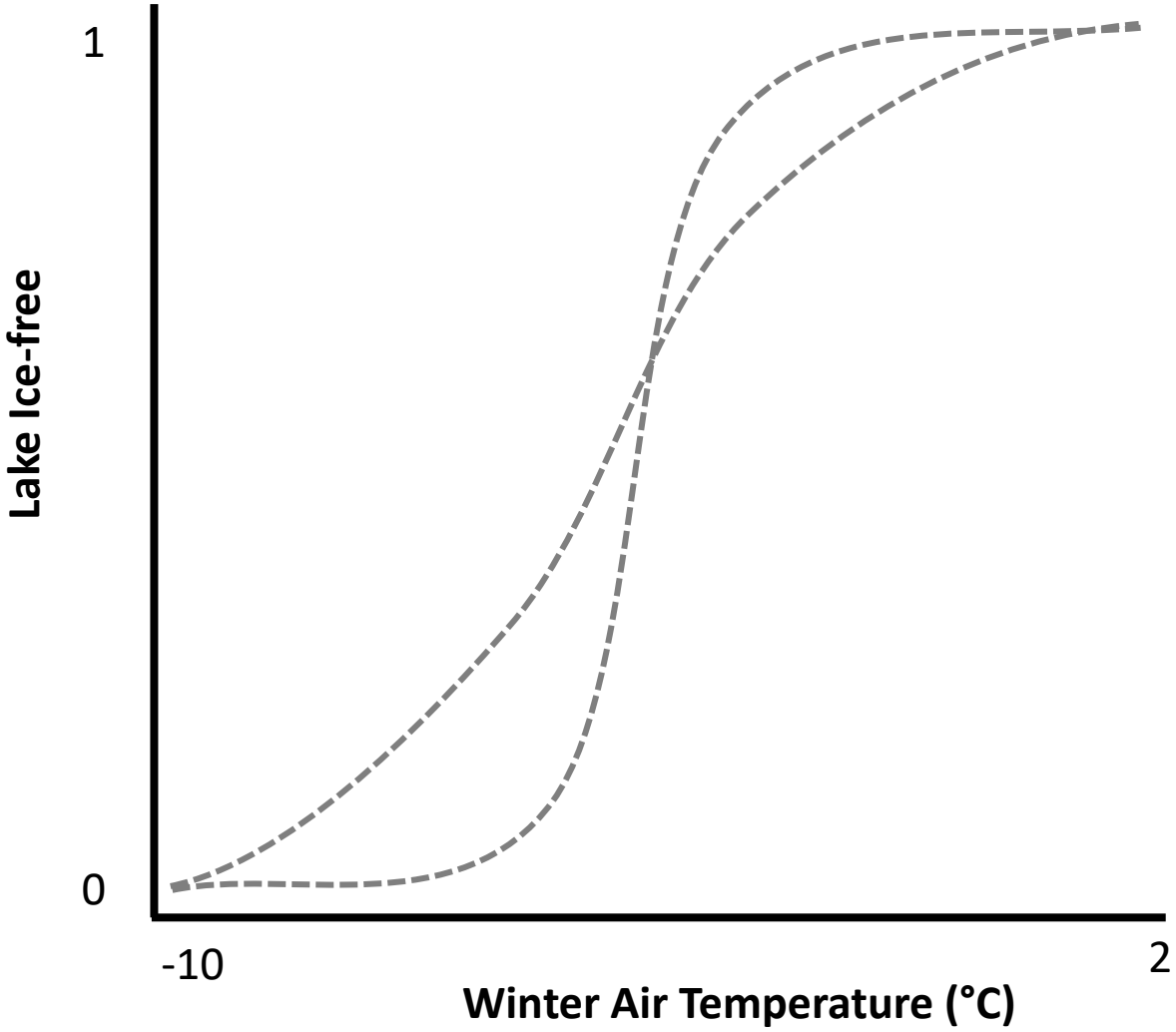


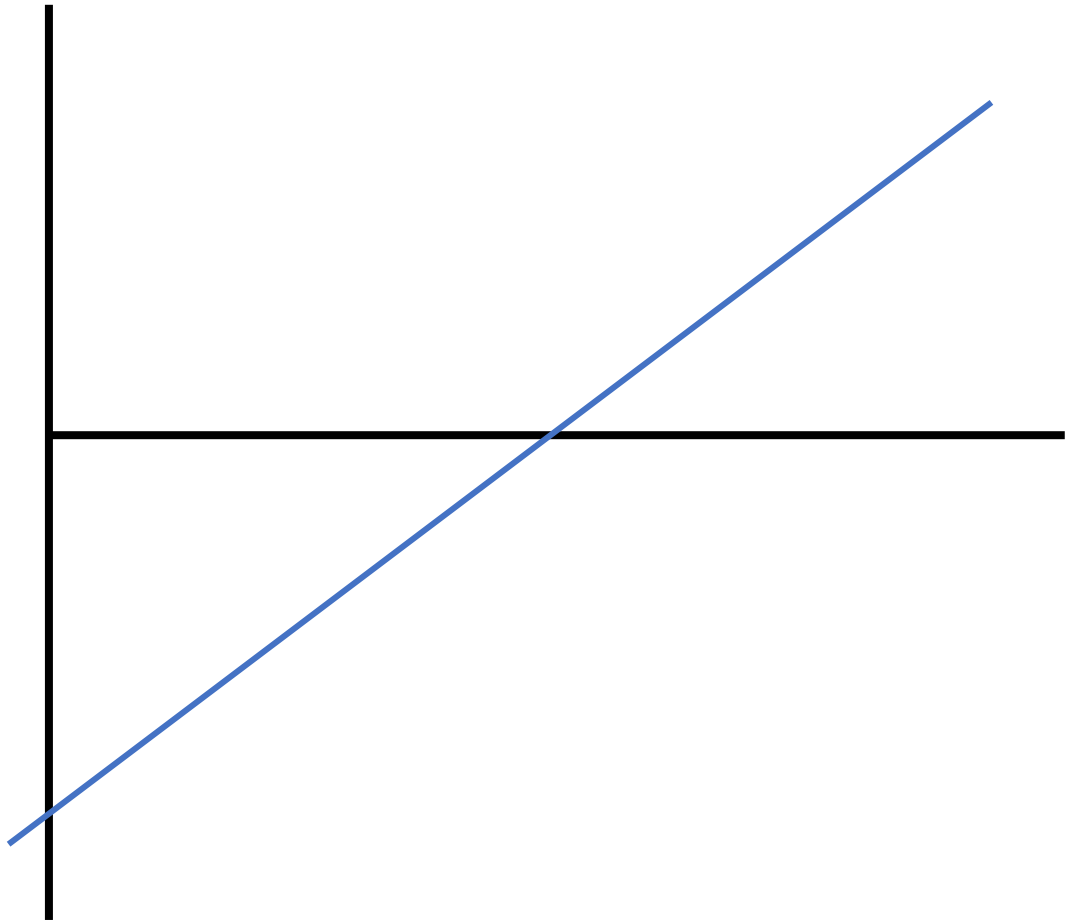
# 1) Fitting the model – maximum likelihood



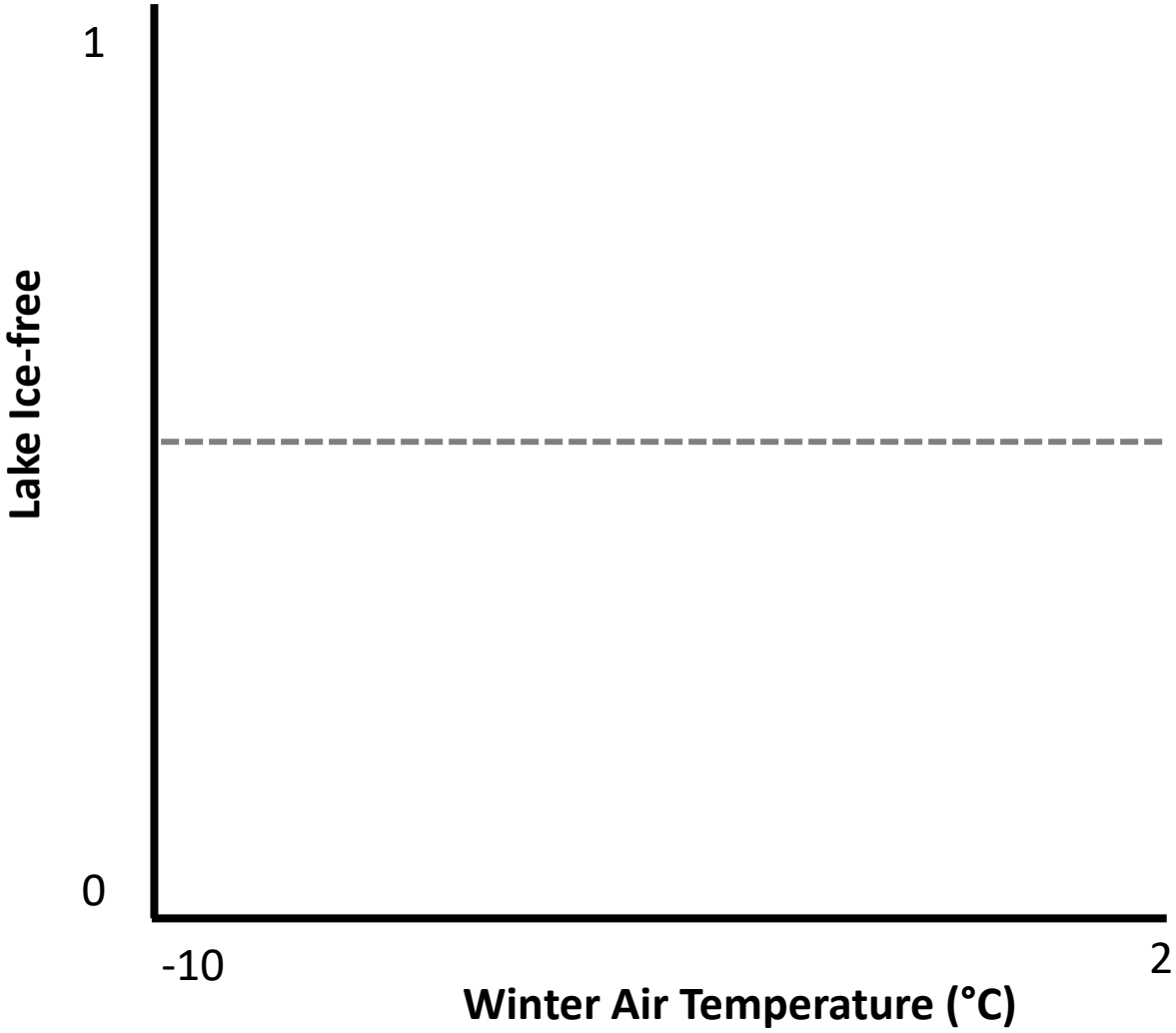



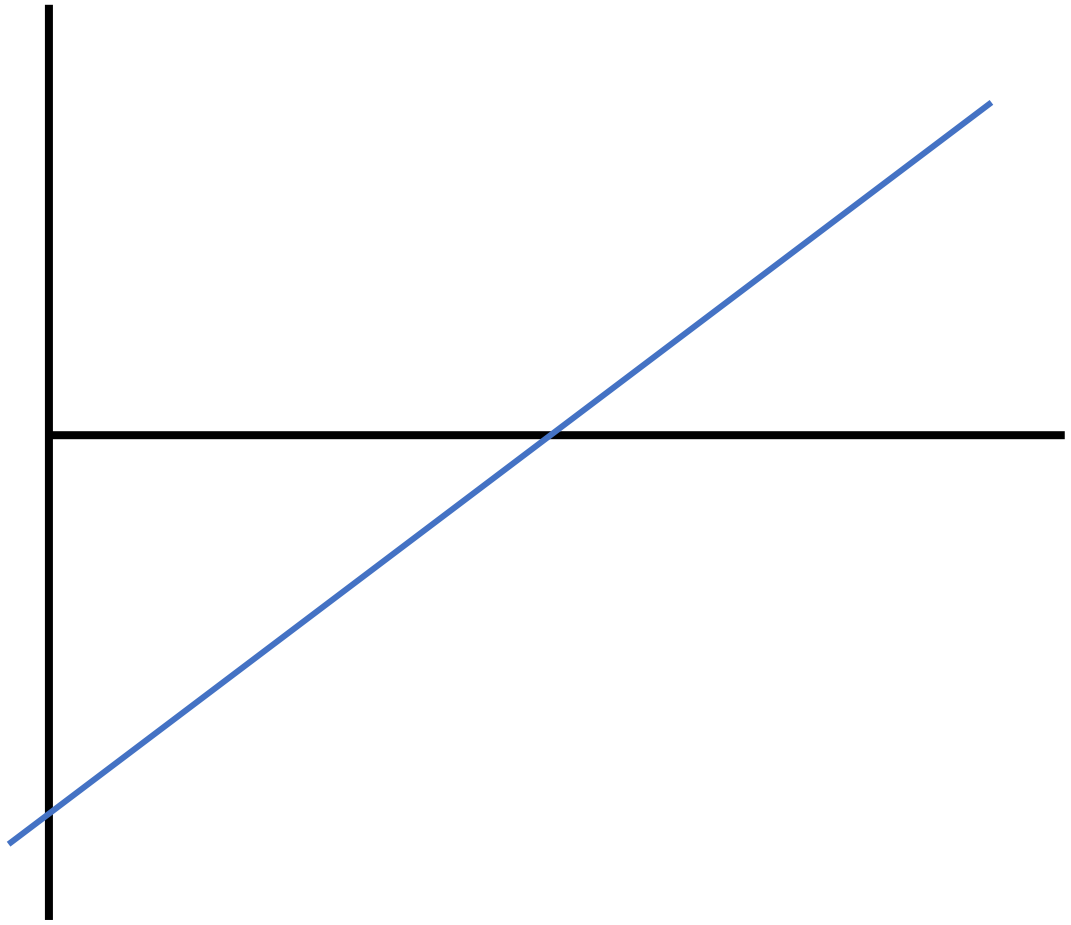
# 1) Fitting the model – maximum likelihood



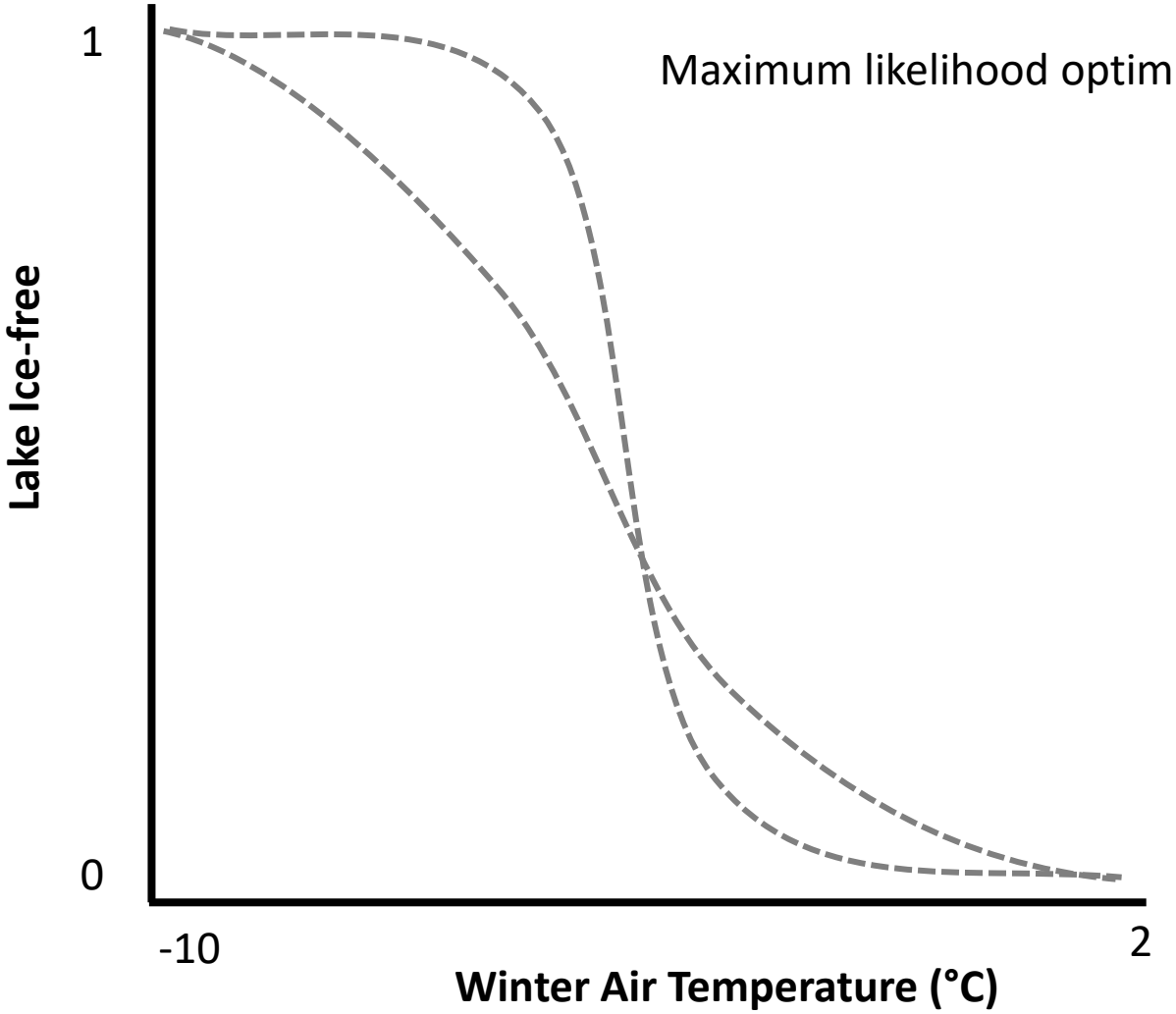



# 1) Fitting the model – maximum likelihood

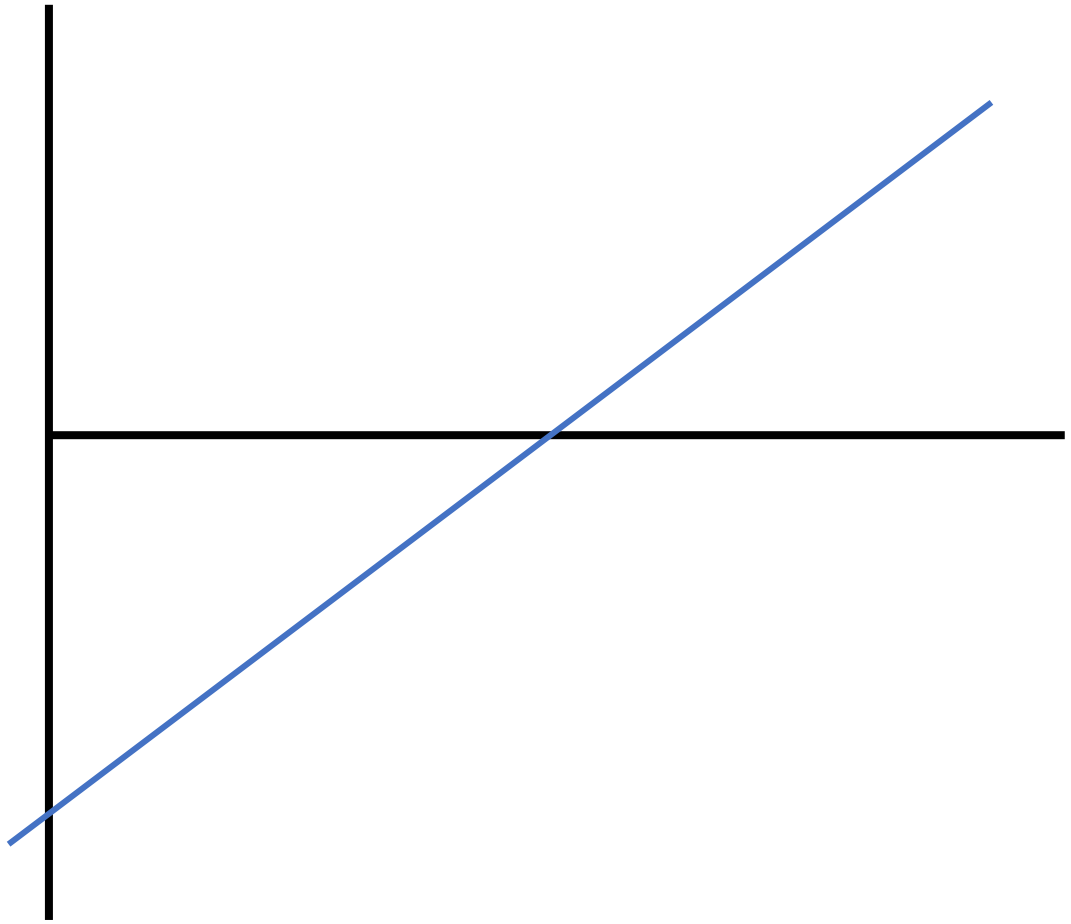


# 1) Fitting the model – maximum likelihood




Maximum likelihood optimizes quickly

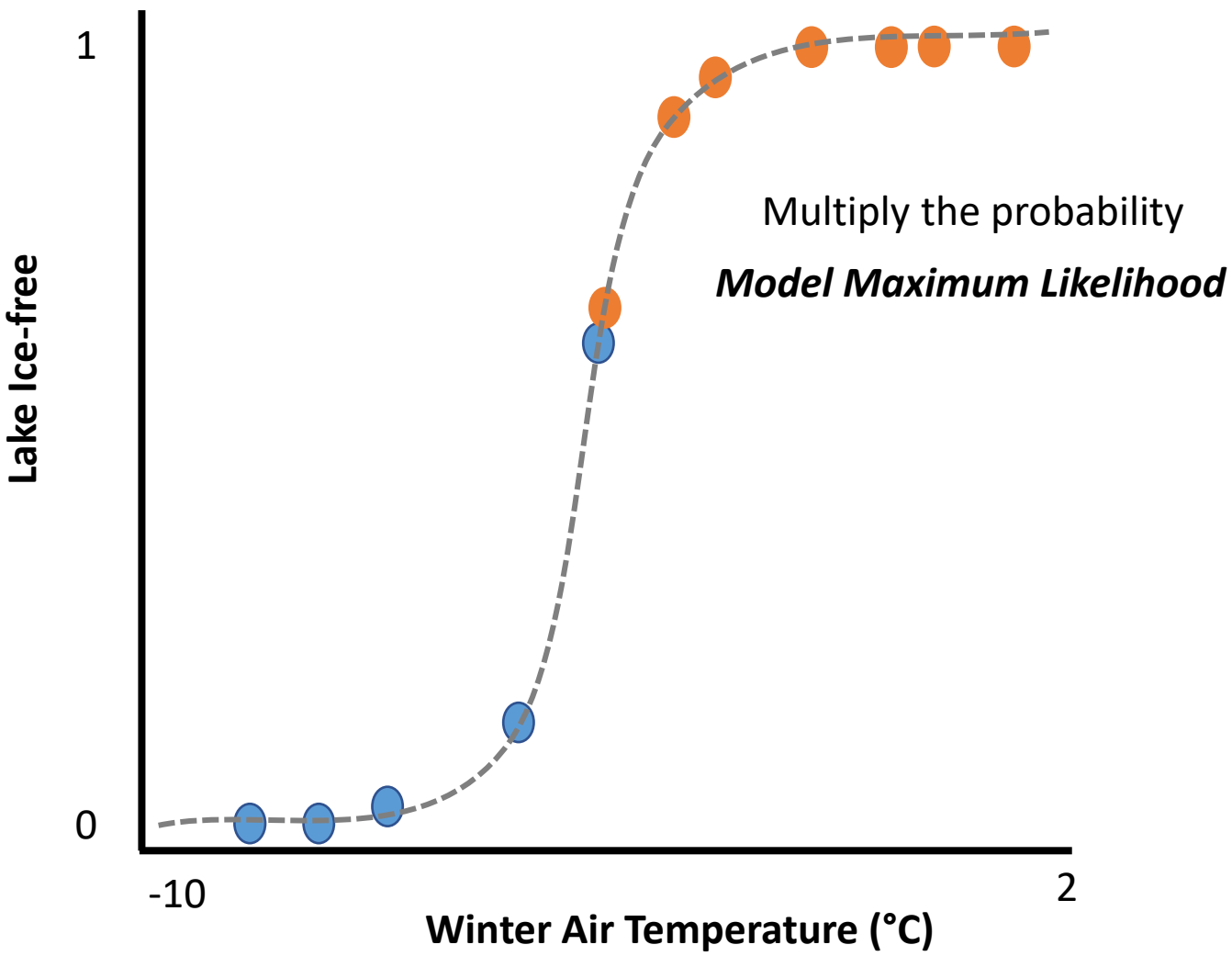


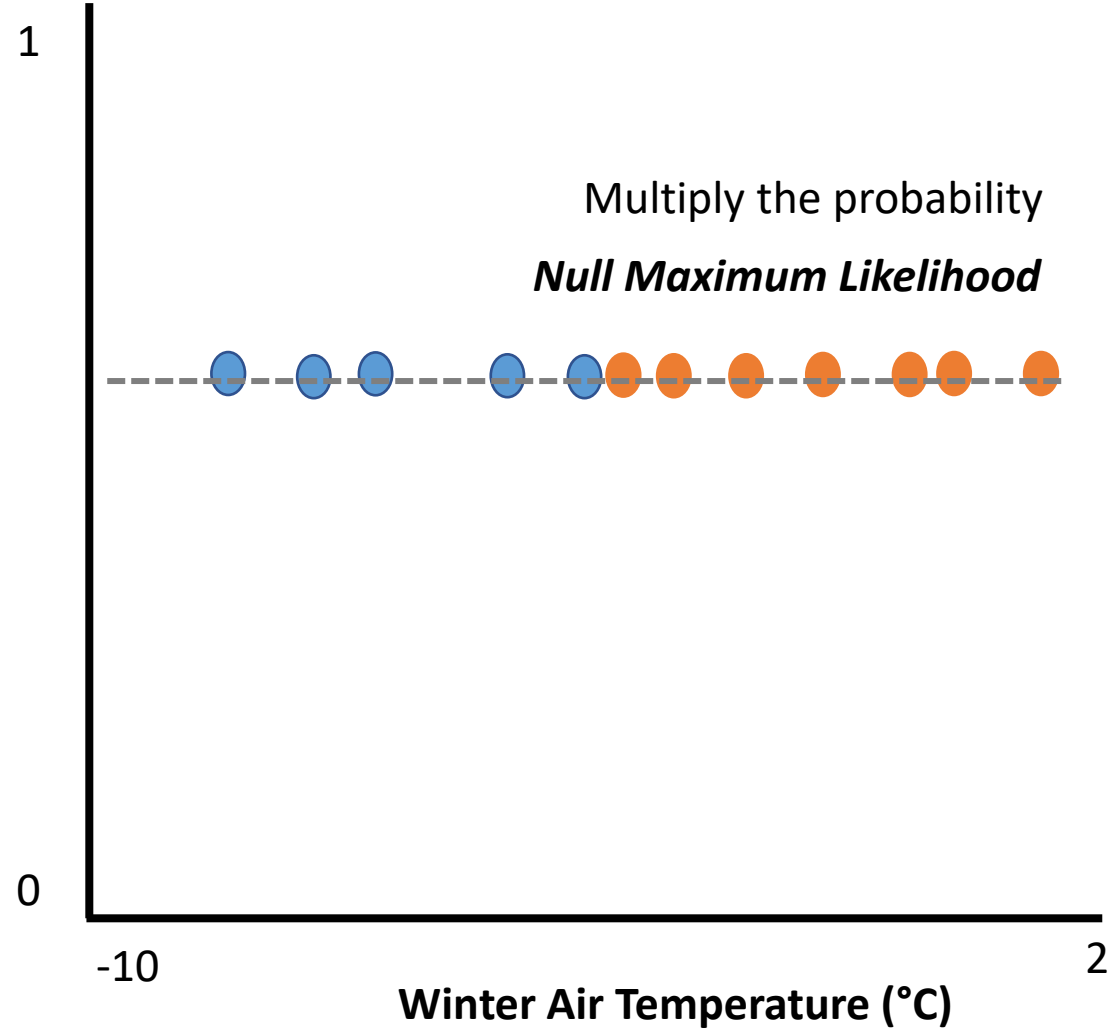
# 2) The outputs – GLMs in R

# 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

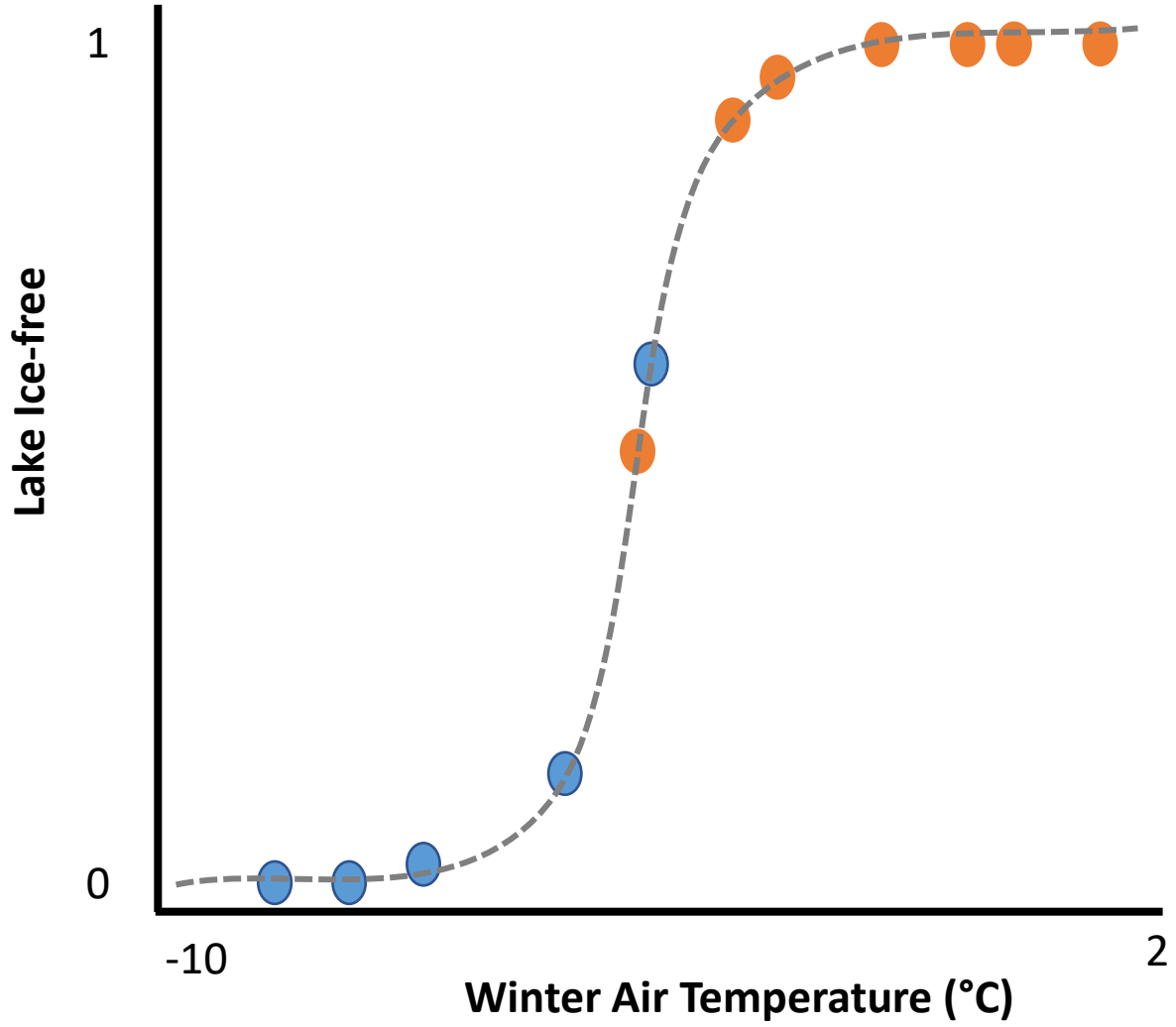
$$\text{PsuedoR}^2 = 1 - \frac{\log(Model)}{\log(Null)}$$

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

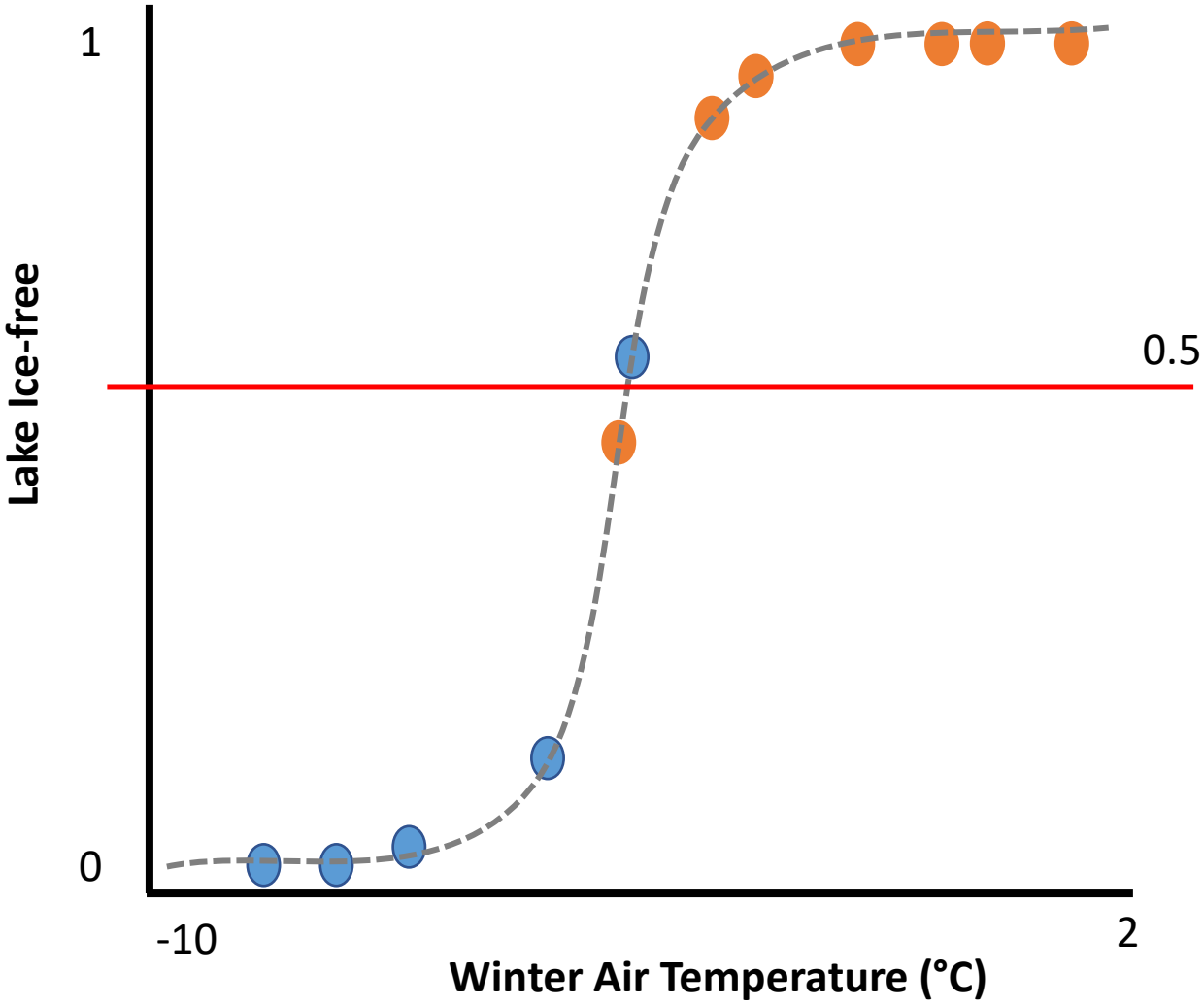
Part B

# Prediction

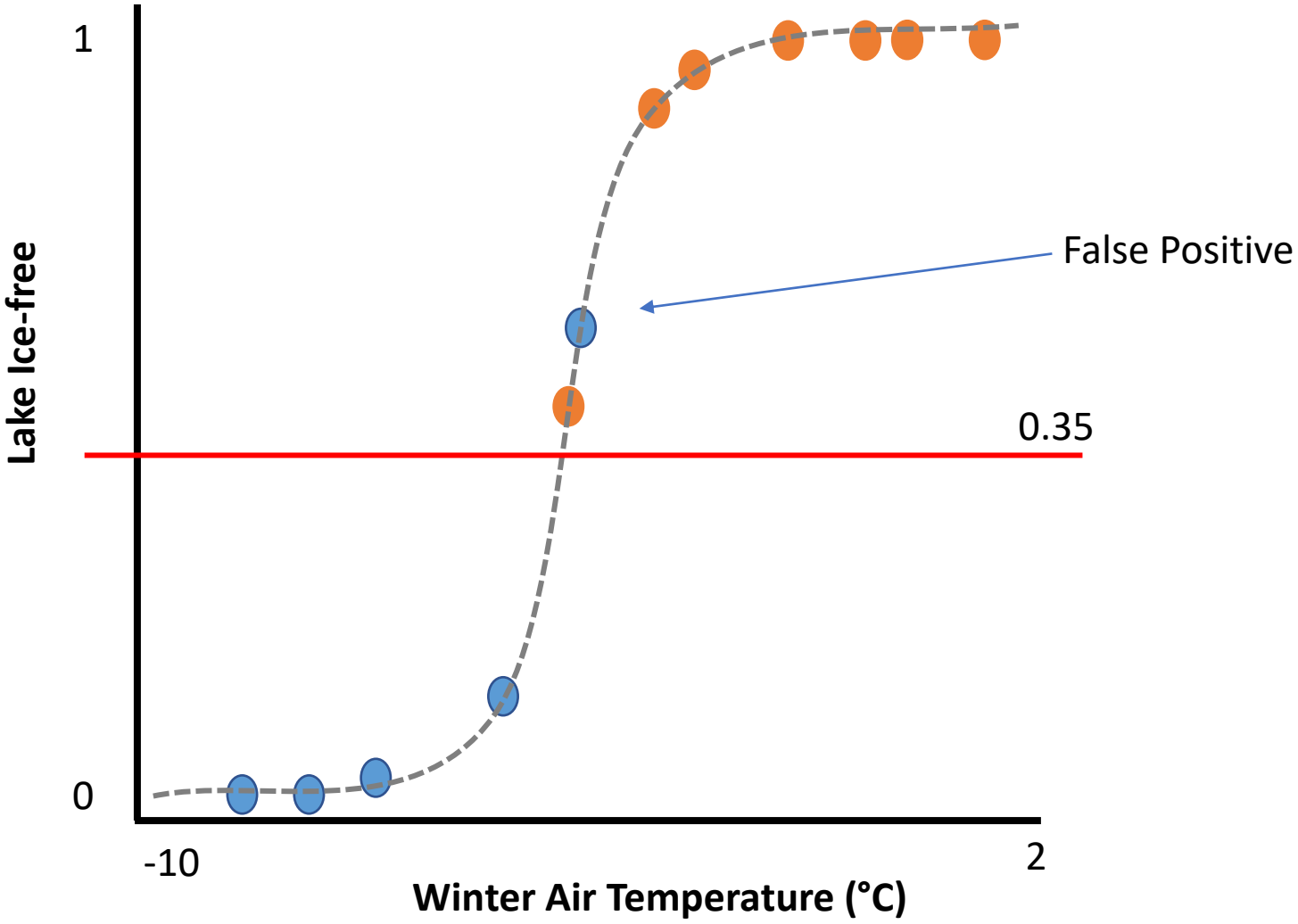
- Models probability
- ..but the outcome is binary
- Need to identify a threshold



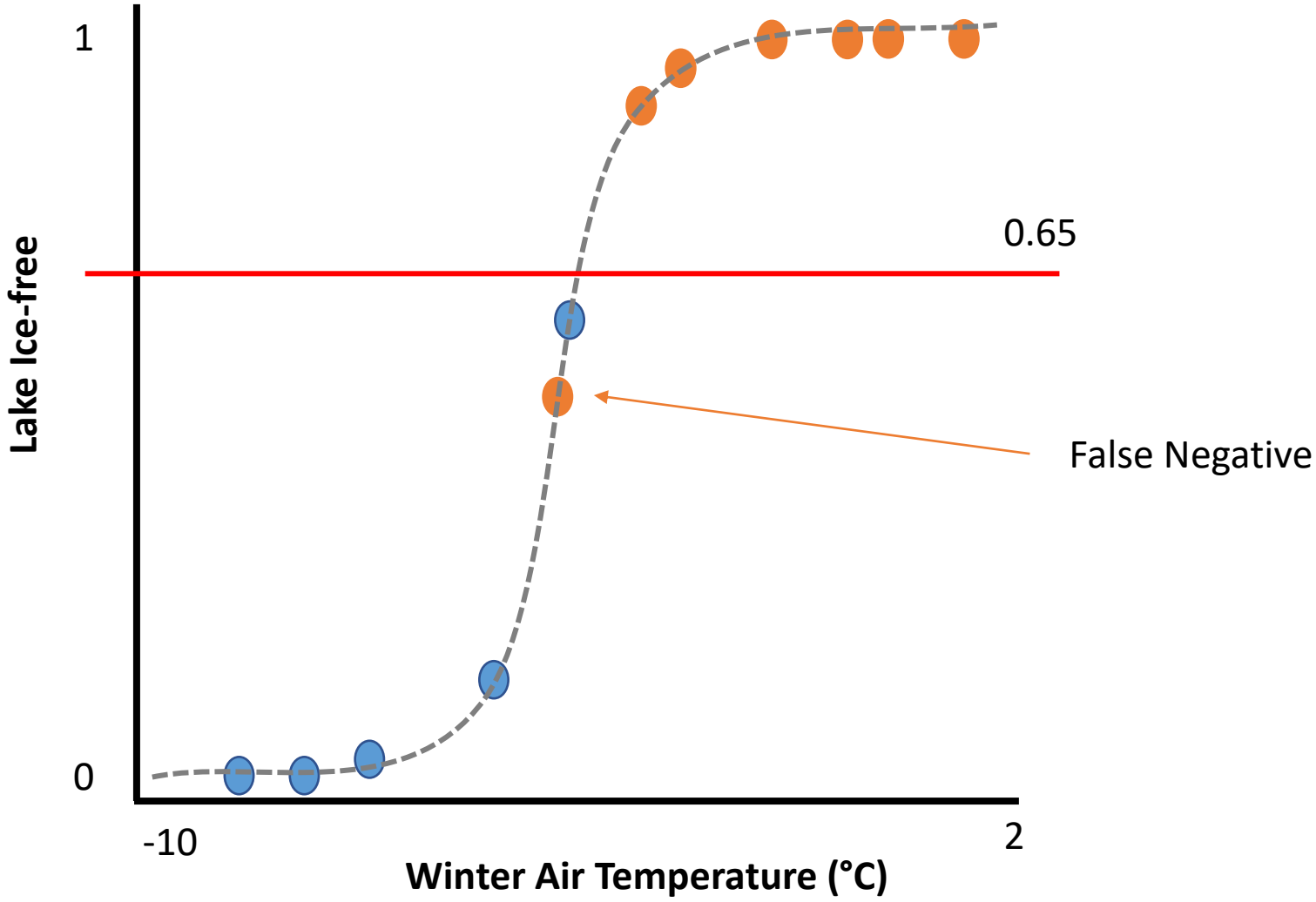
# Prediction



# Prediction


1
0
Lake Ice-free
False Positive
0.35
-10
2
Winter Air Temperature (°C)

# Prediction

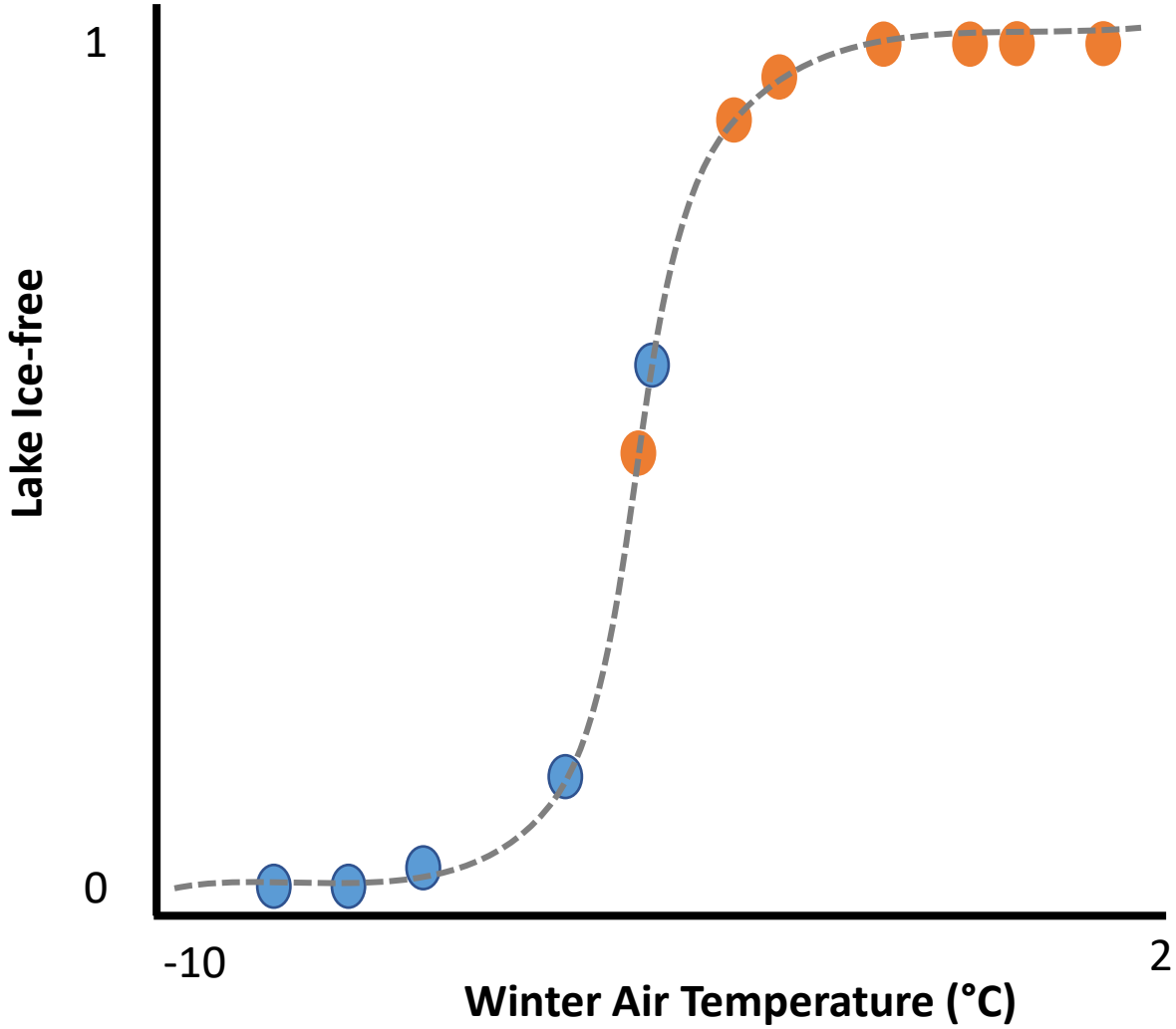

1
0.65
0
Lake Ice-free
False Negative
-10
2
Winter Air Temperature (°C)

# Question

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

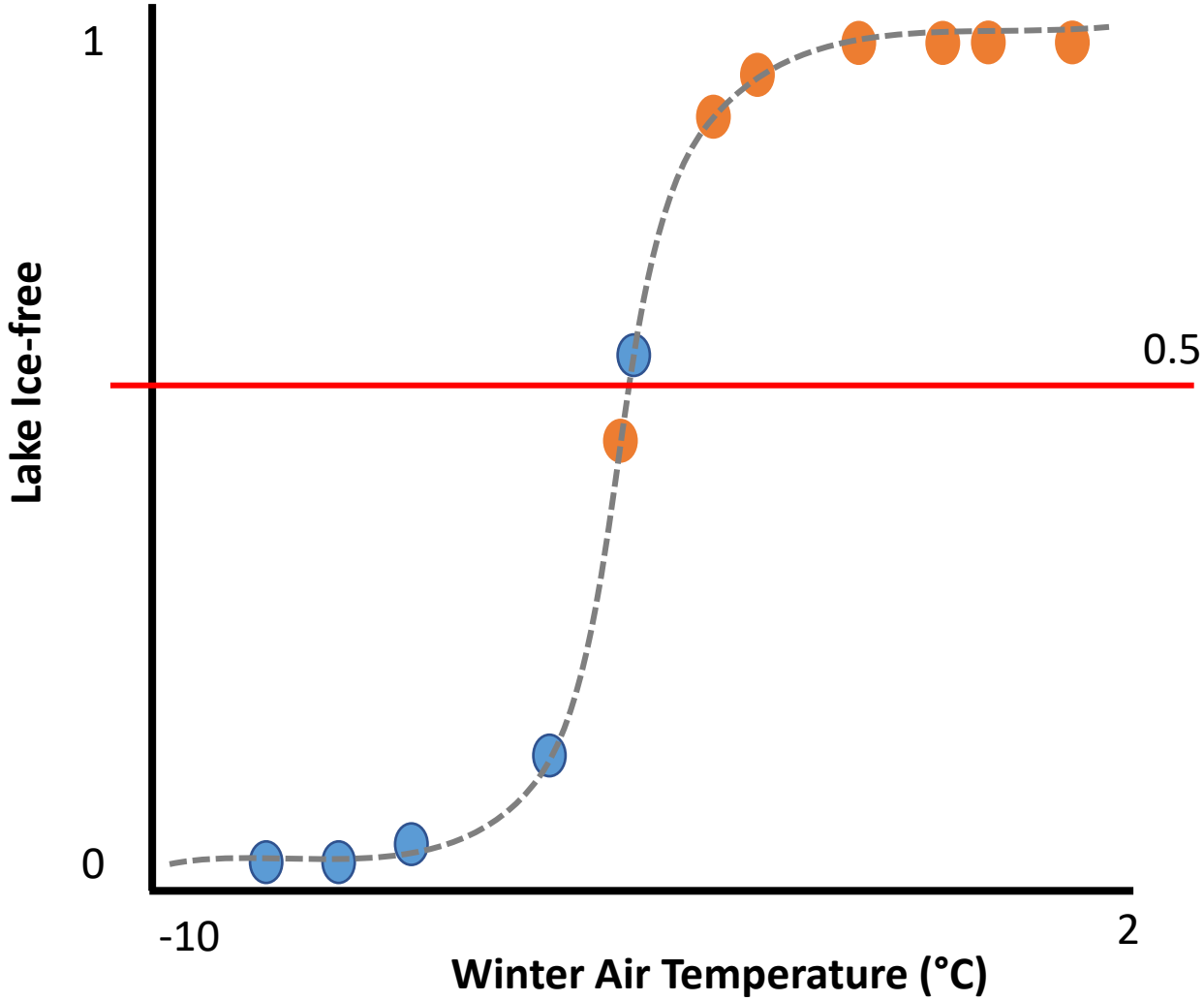
a) Predicting infection of a disease
b) Understanding climate effects on lake ice
c) Predicting the probability of a species occurring
d) A and C
e) All the above

# Prediction – confusion matrix



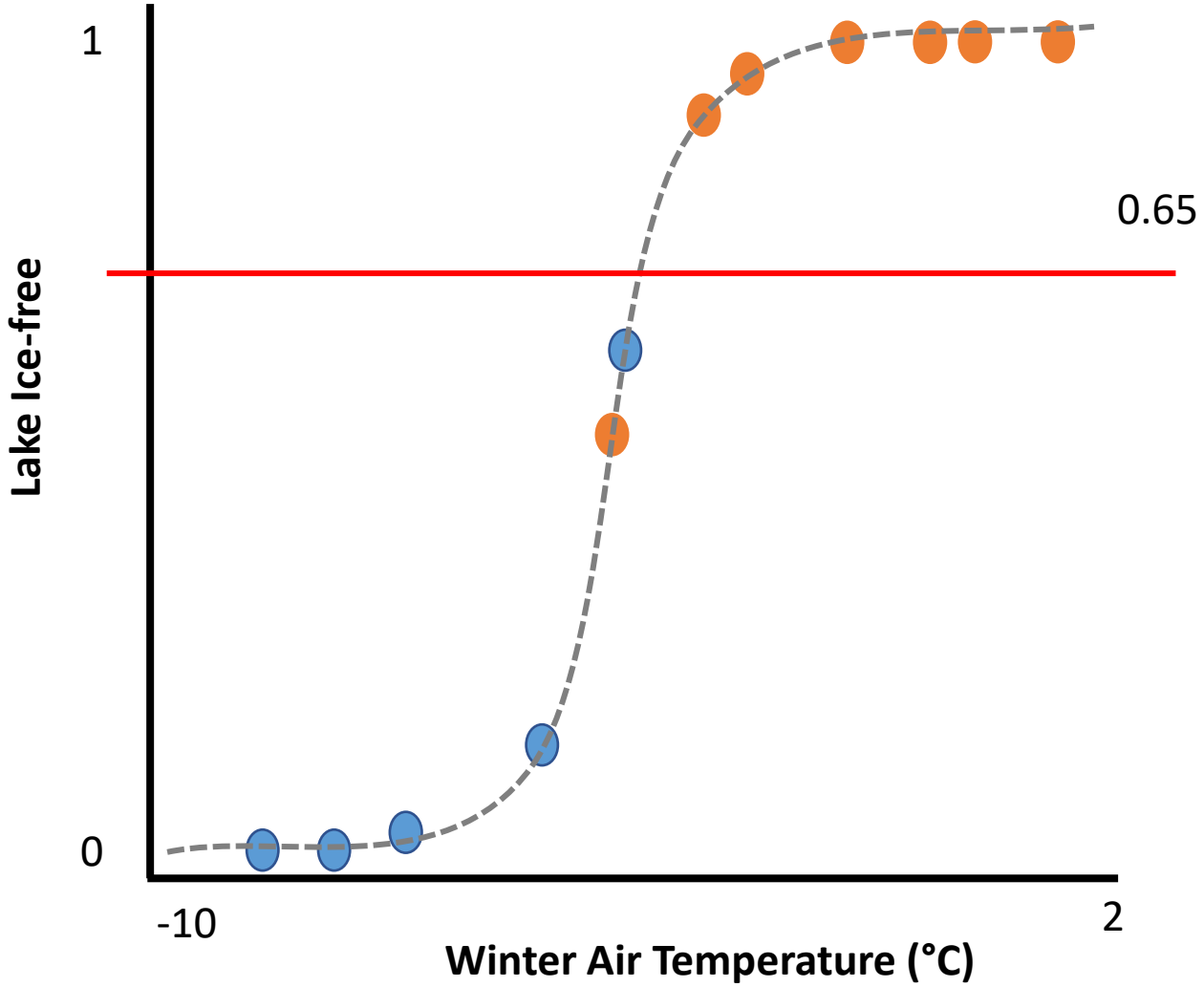

| | **Actual Positive** | **Actual Negative** |
|---|---|---|
| **Predicted Positive** | True positive | False positive |
| **Predicted Negative** | False negative | True negative |

# Prediction – confusion matrix



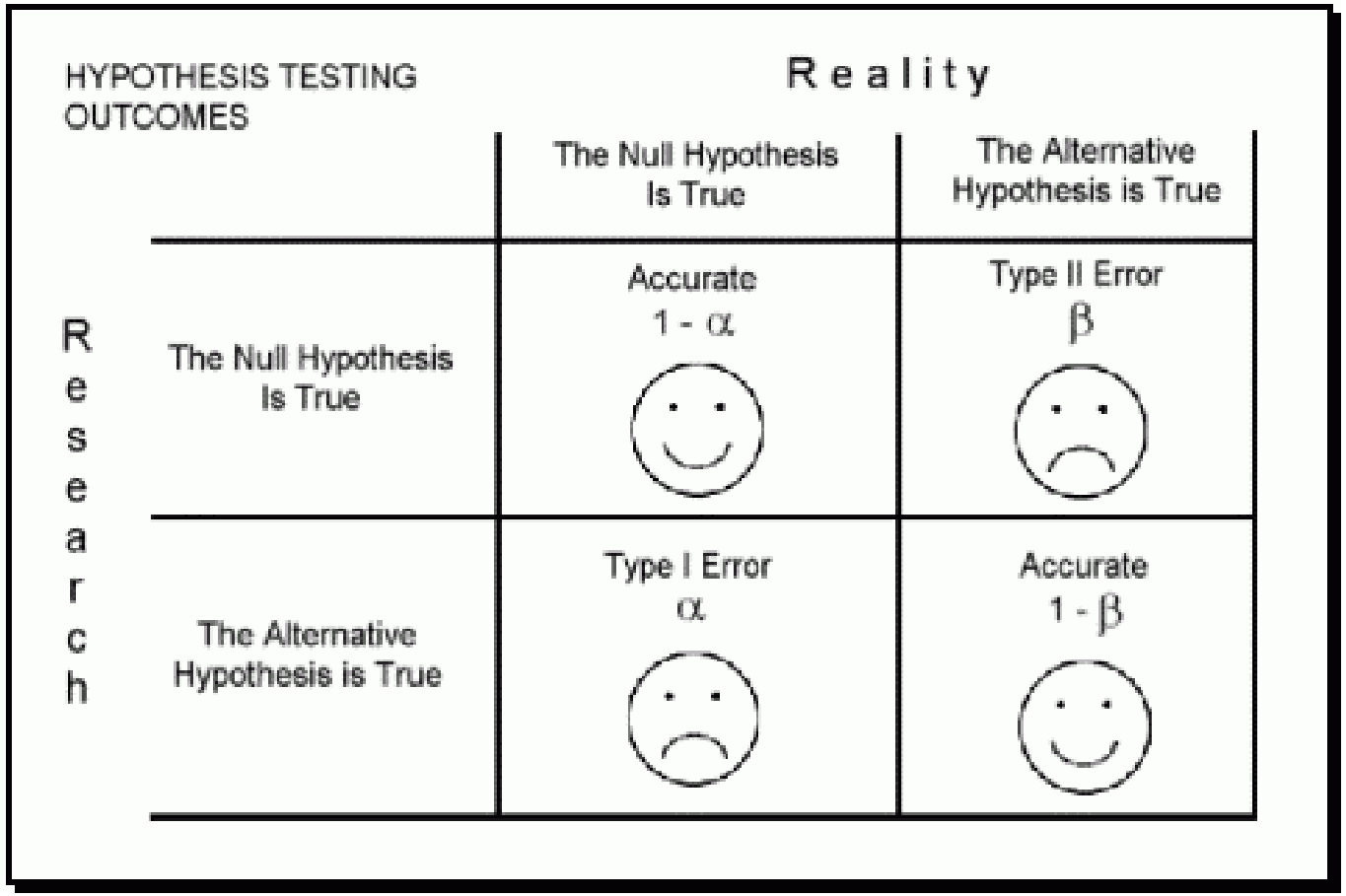

| | Actual Positive | Actual Negative |
|---|---|---|
| **Predicted Positive** | | |
| **Predicted Negative** | | |

# Prediction – confusion matrix




| | Actual Positive | Actual Negative |
|---|---|---|
| **Predicted Positive** | | |
| **Predicted Negative** | | |

# Prediction – confusion matrix




https://blog.datasciencedojo.com/

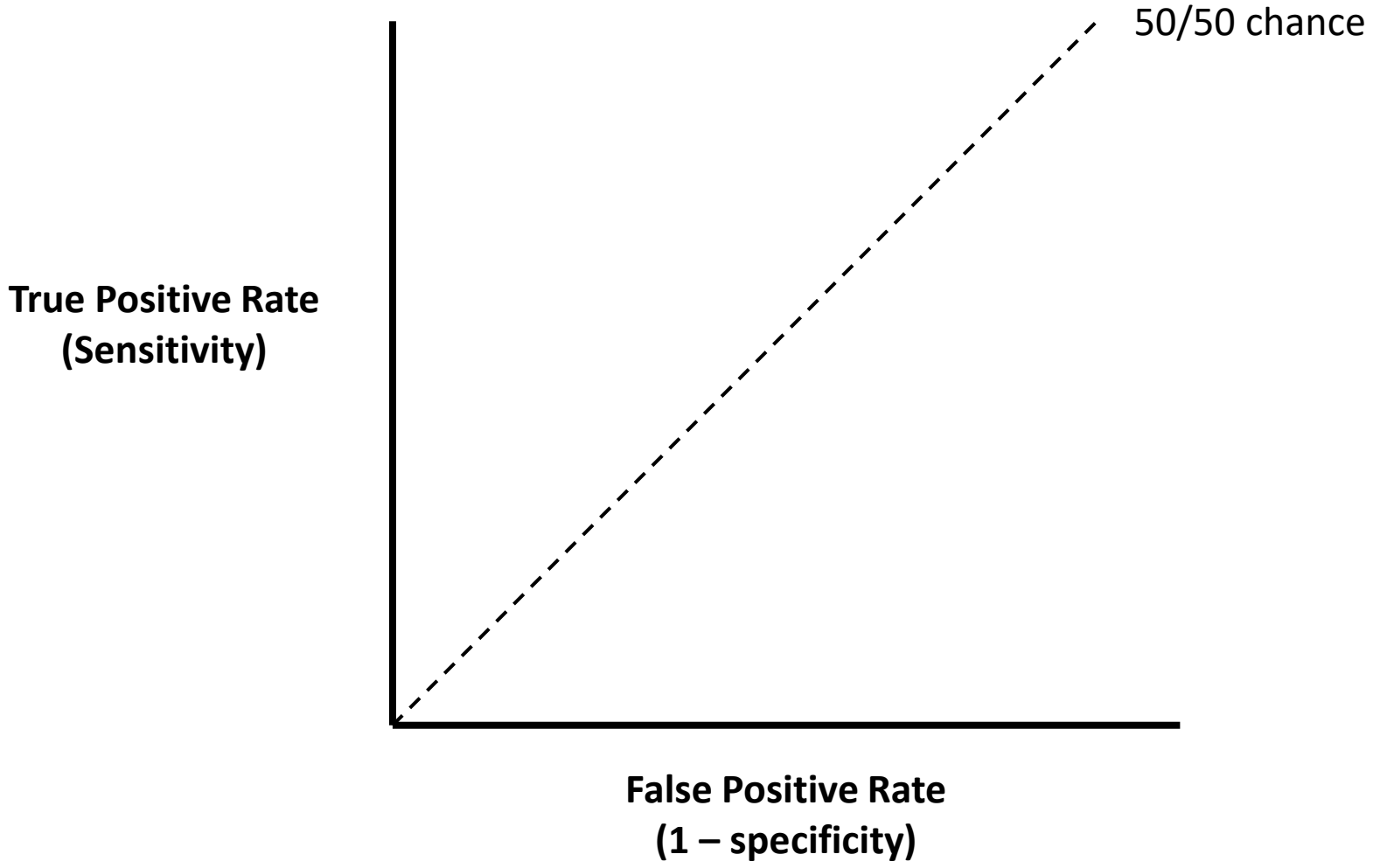
# Optimizing the threshold

Many different methods

Receiver operator characteristic (ROC) often used to pick threshold

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

# Optimizing the threshold - ROC



# 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**

# Optimizing the threshold - ROC

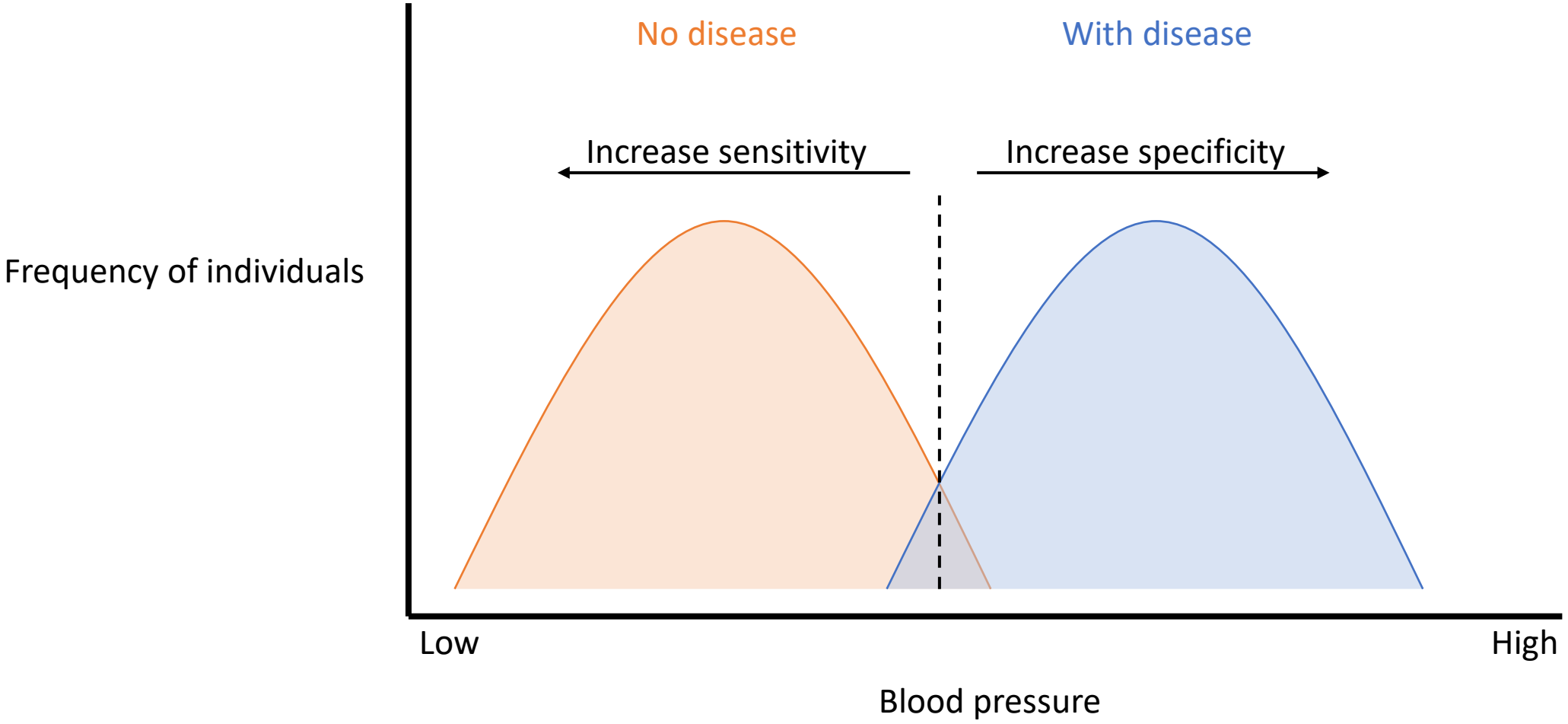
| | **Actual Positive** | **Actual Negative** |
|---|---|---|
| **Predicted Positive** | True positive | False positive |
| **Predicted Negative** | False negative | True negative |

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

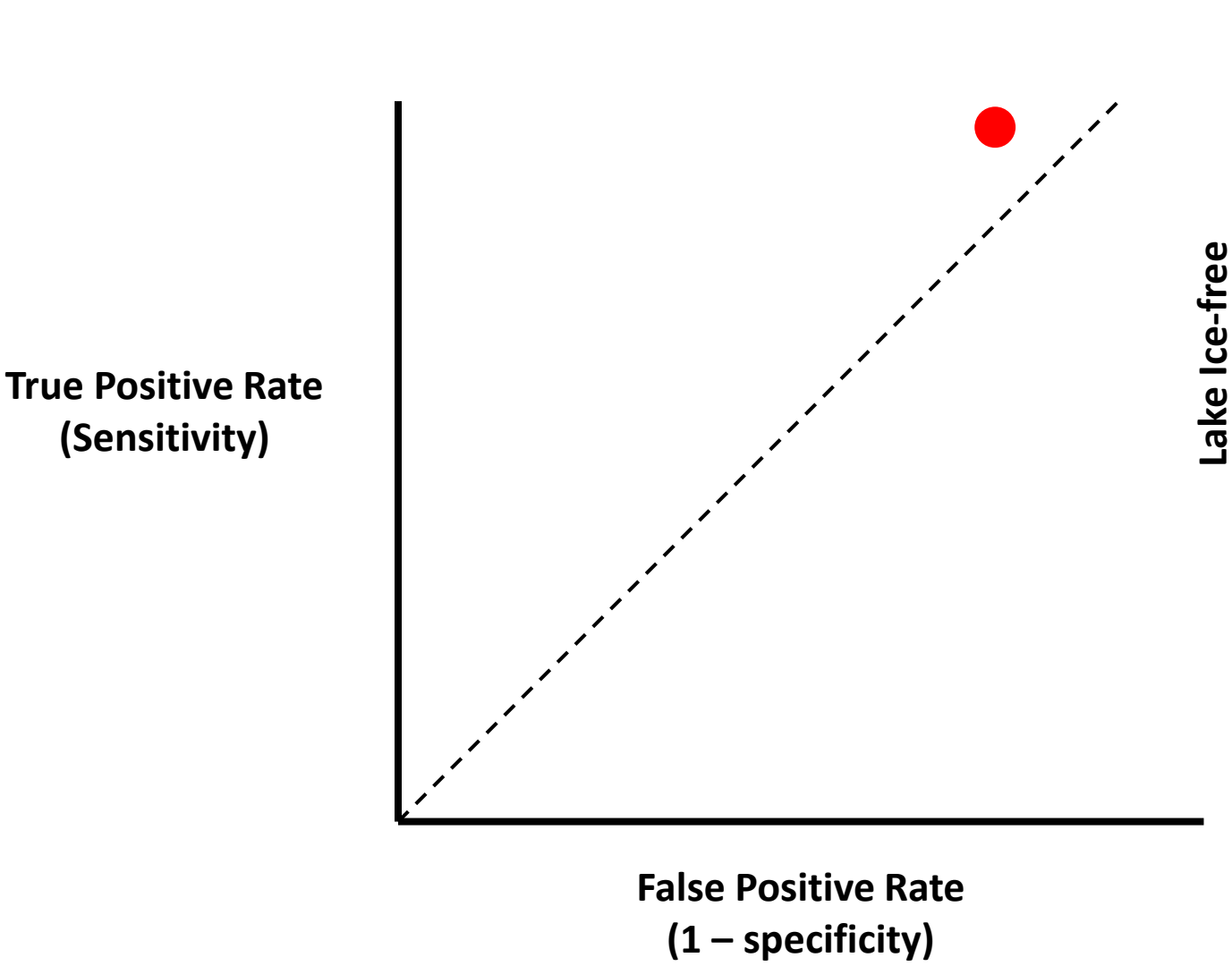
$$\text{False positive rate} = \frac{\text{True positives}}{\text{False positives + True negatives}}$$

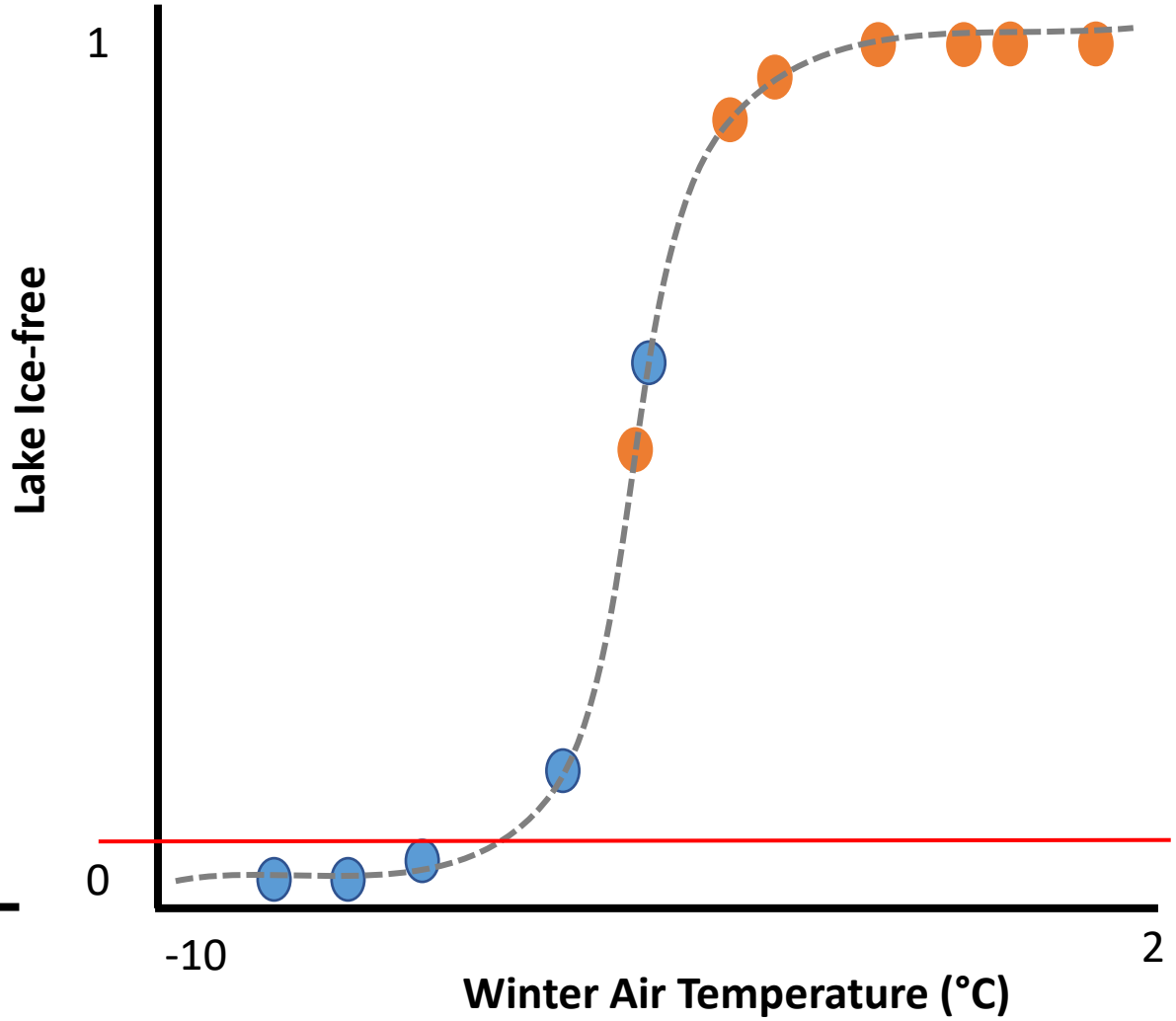
**Specificity**

# Sensitivity vs. specificity

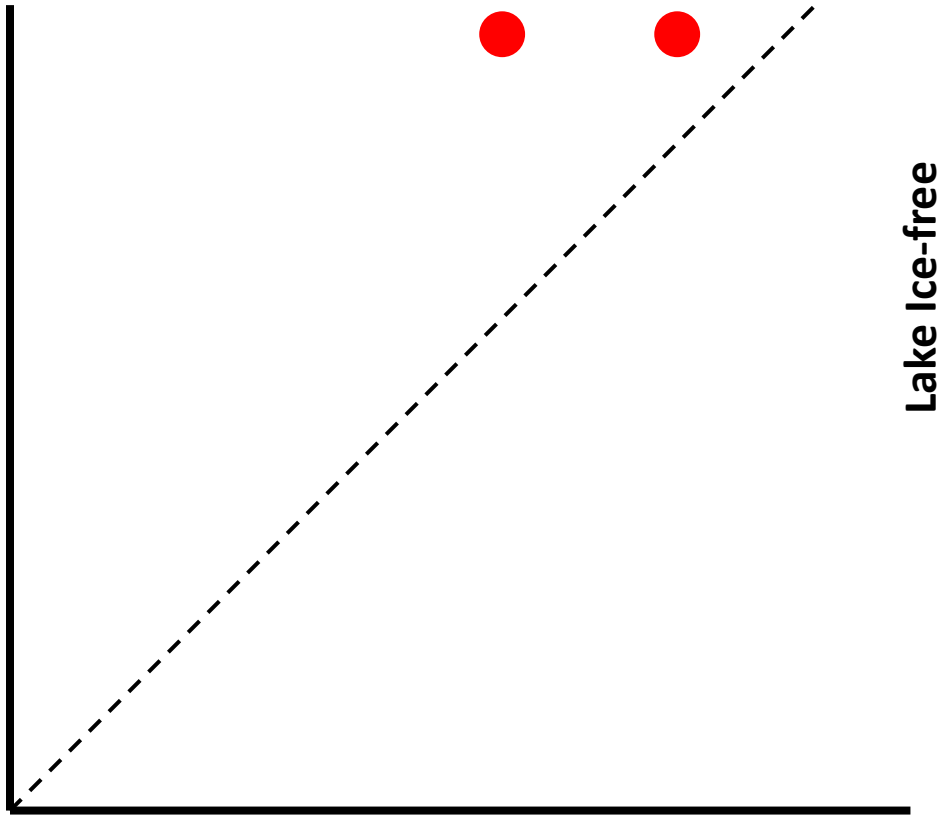


# Optimizing the threshold - ROC



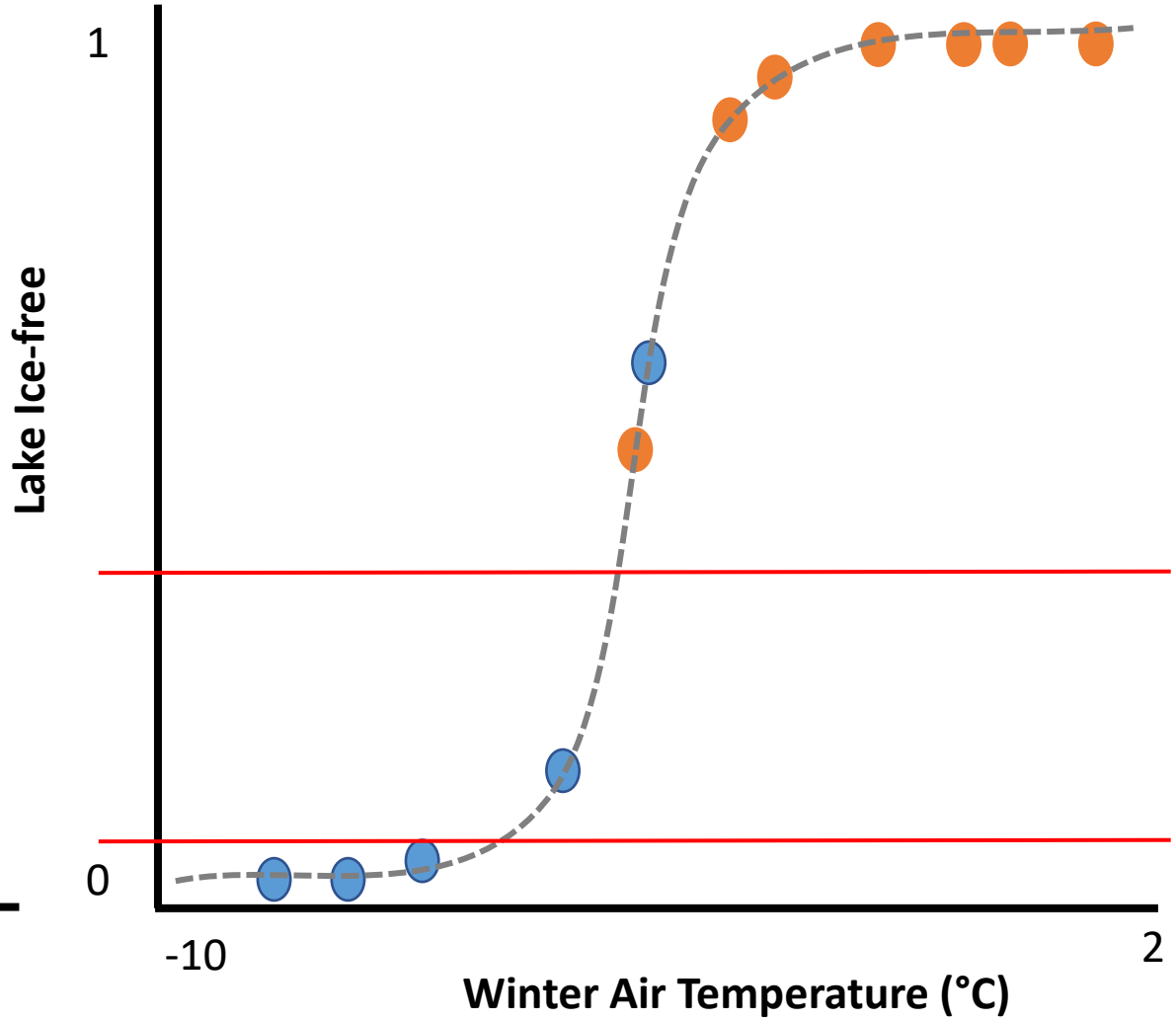



# Optimizing the threshold - ROC

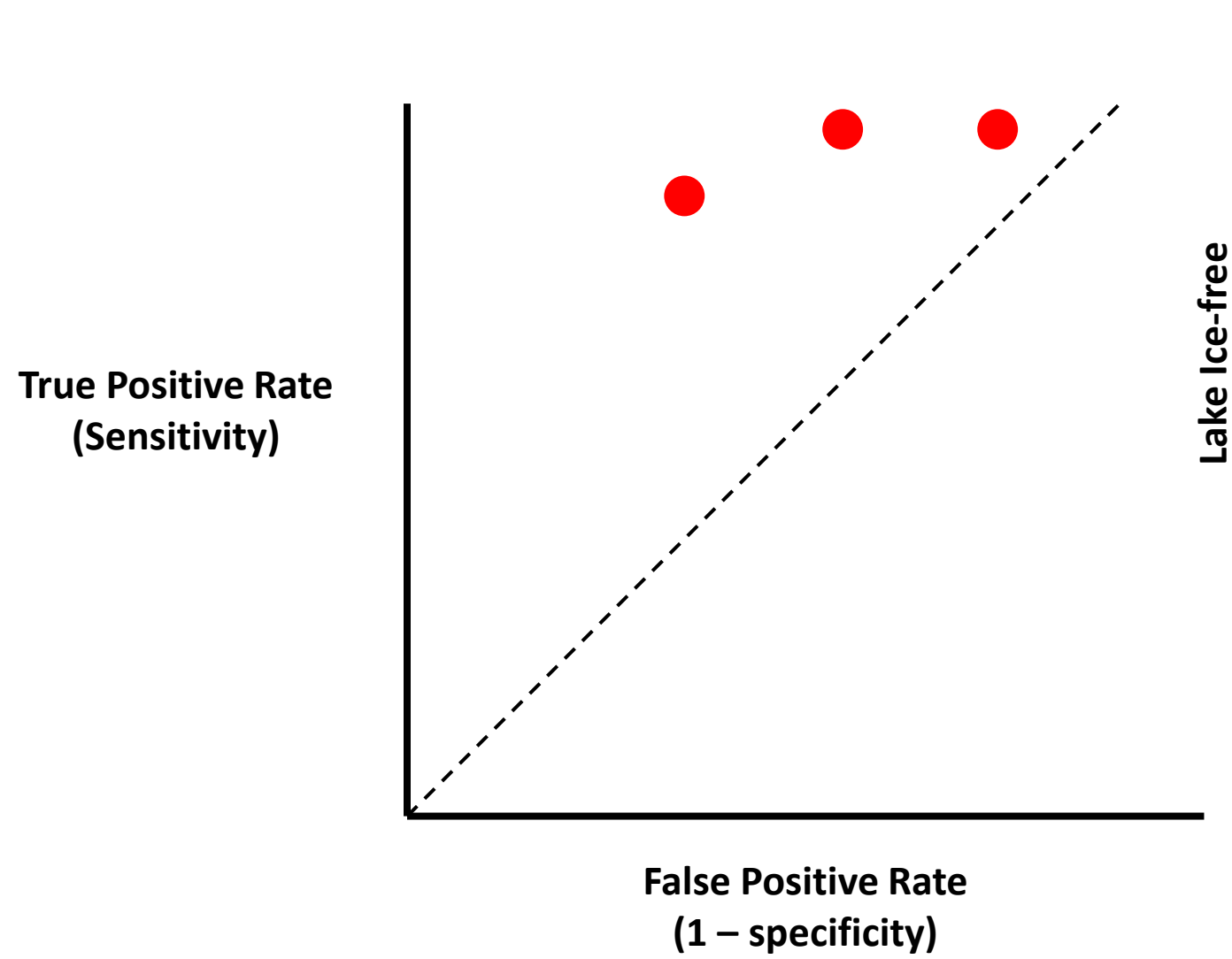


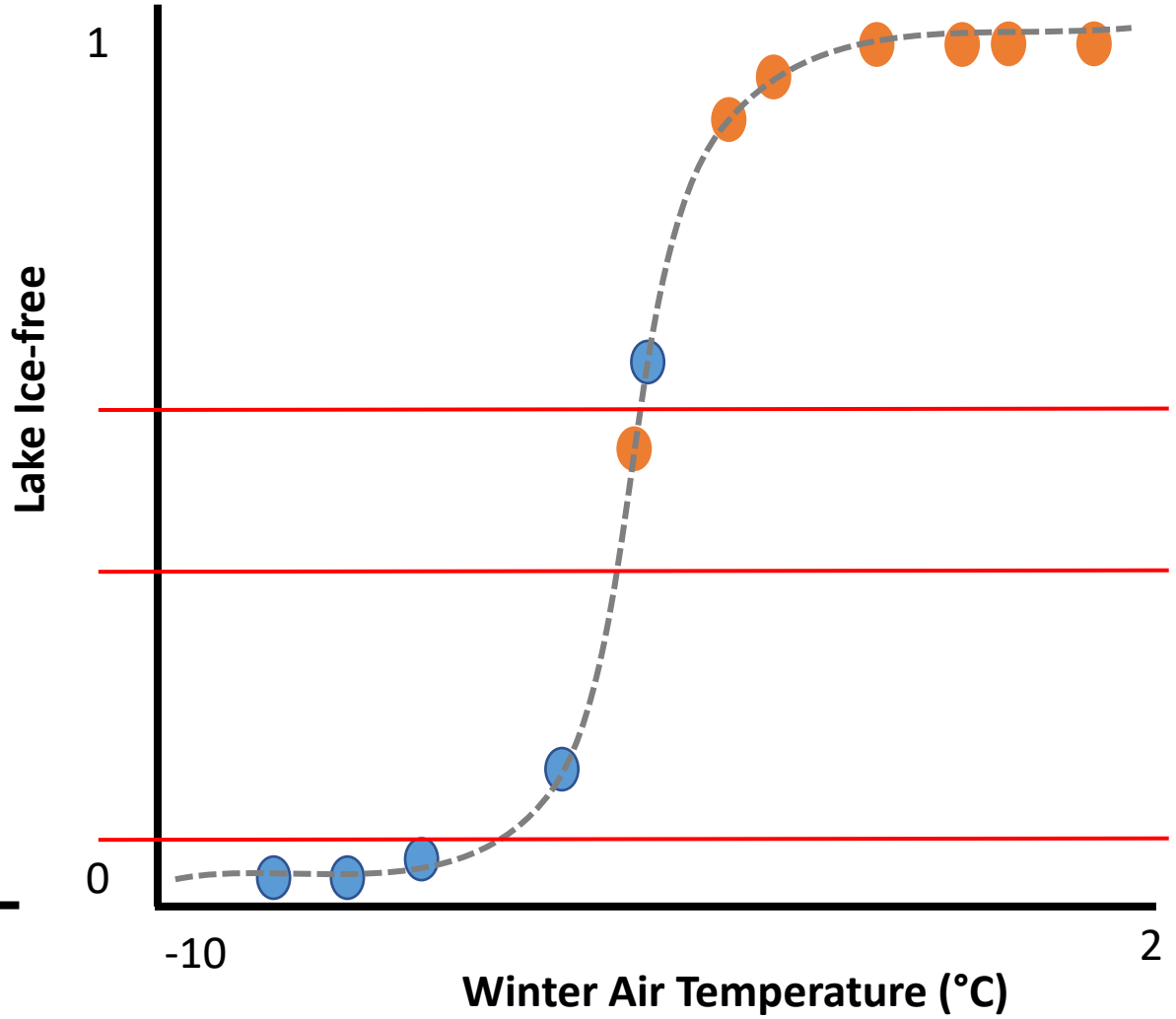
True Positive Rate
(Sensitivity)

False Positive Rate
(1 – specificity)

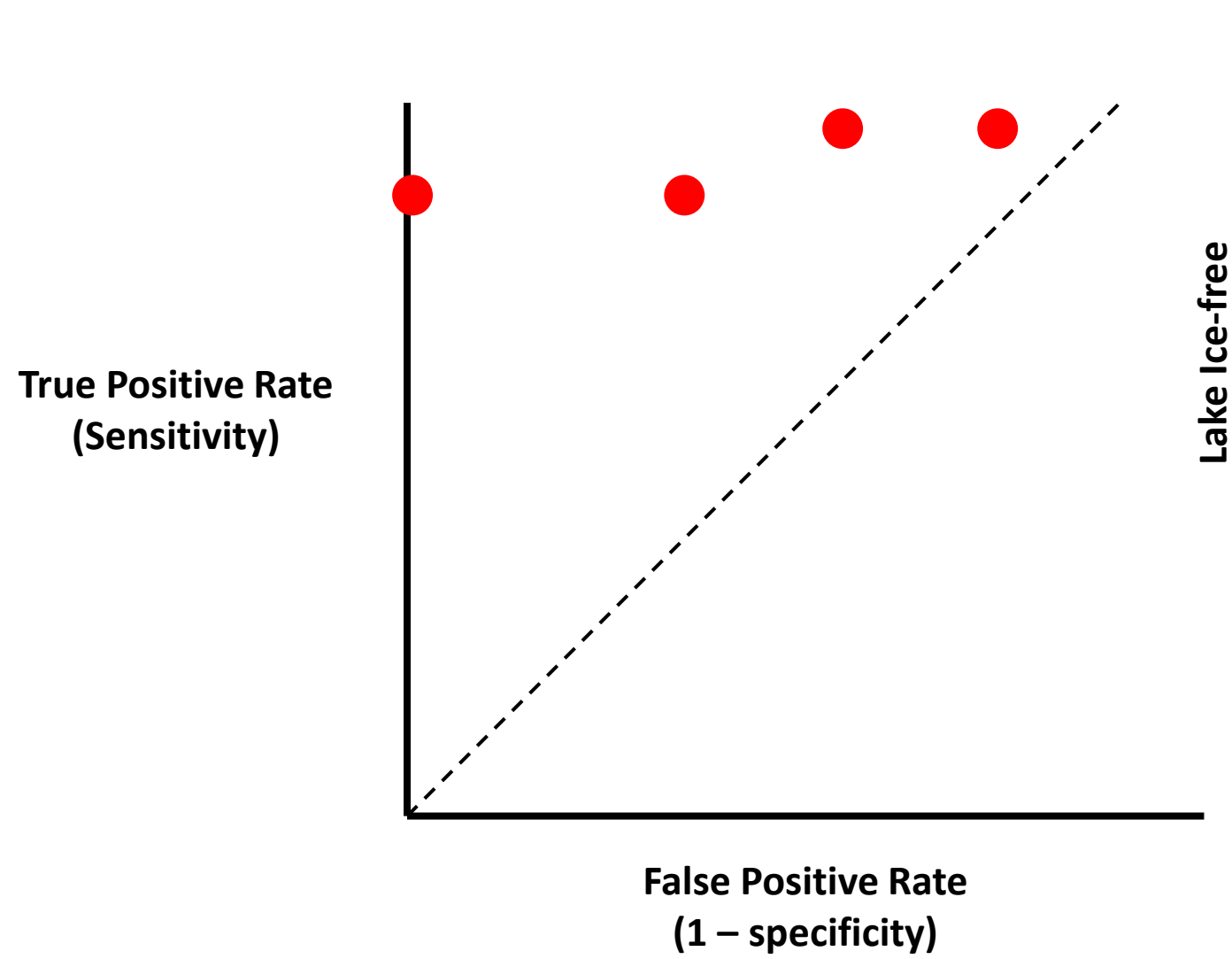


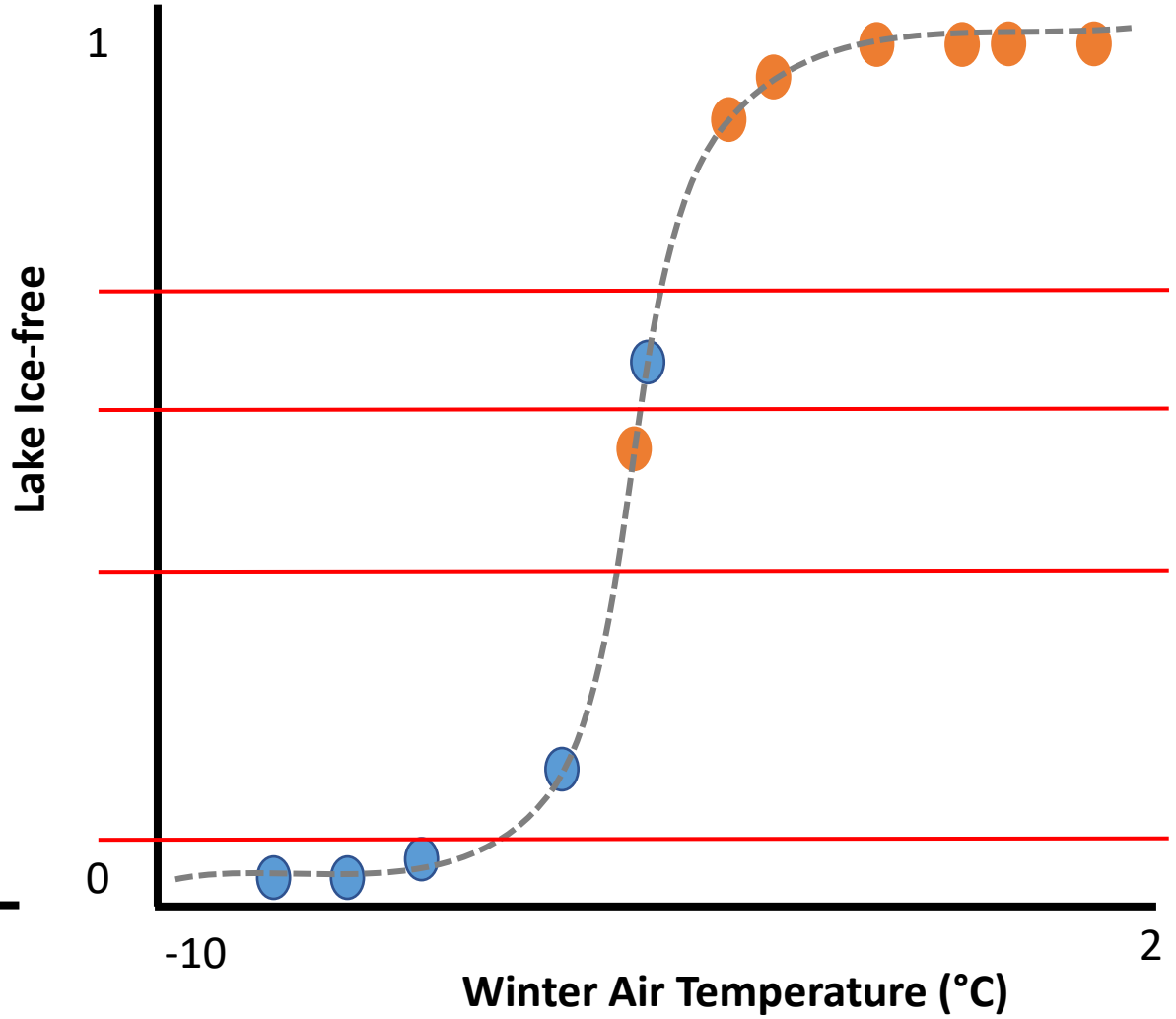
# Optimizing the threshold - ROC



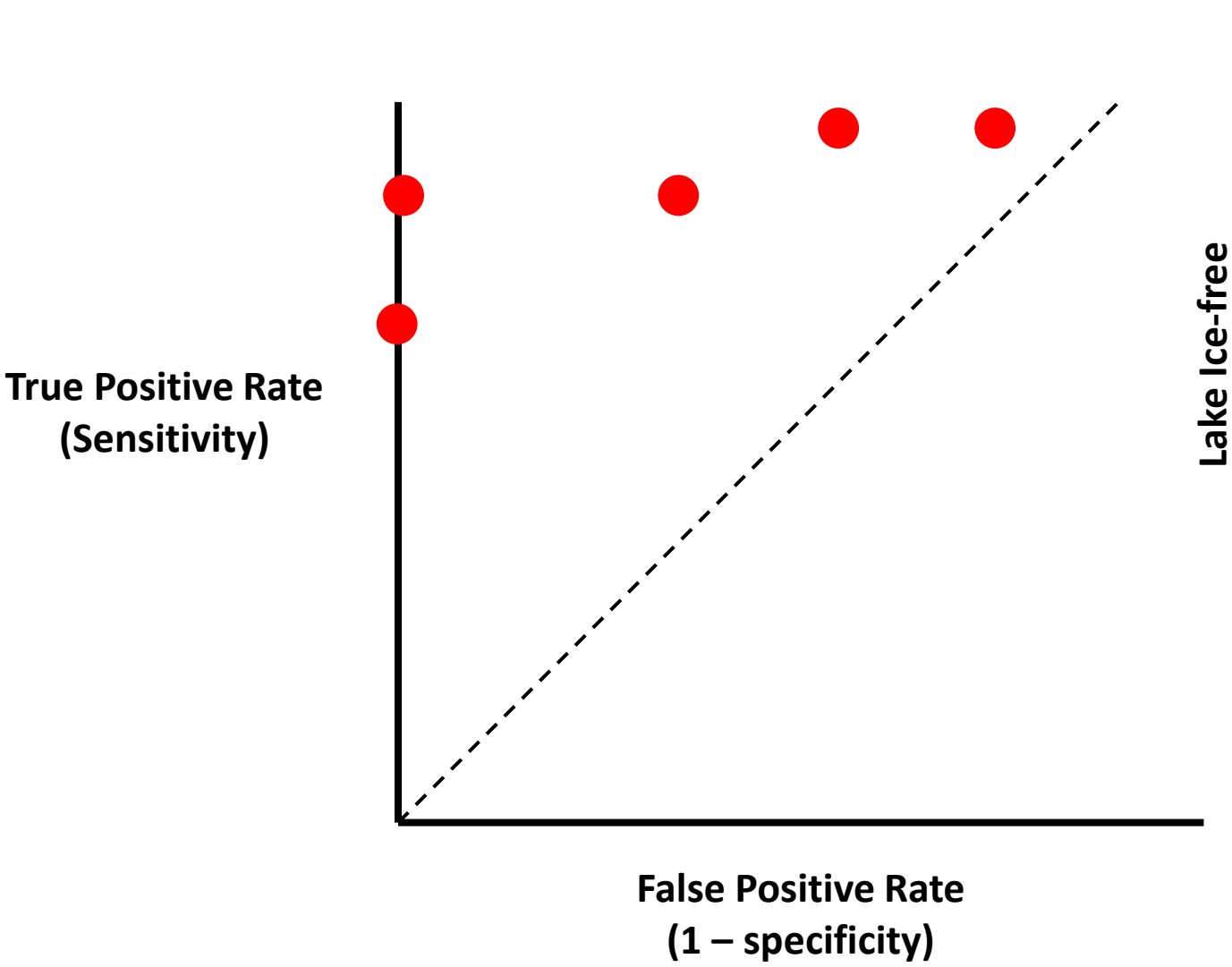



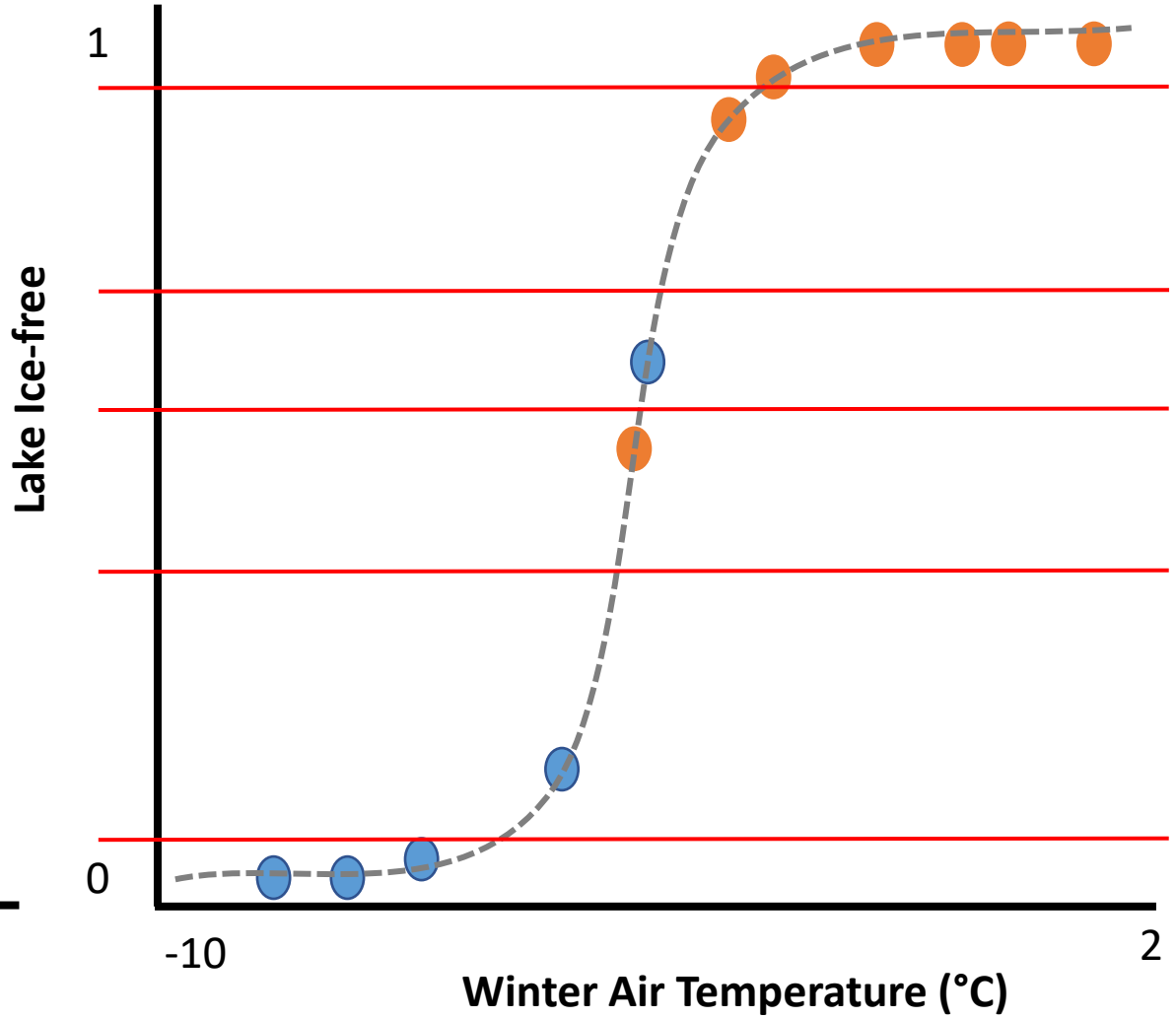
# Optimizing the threshold - ROC



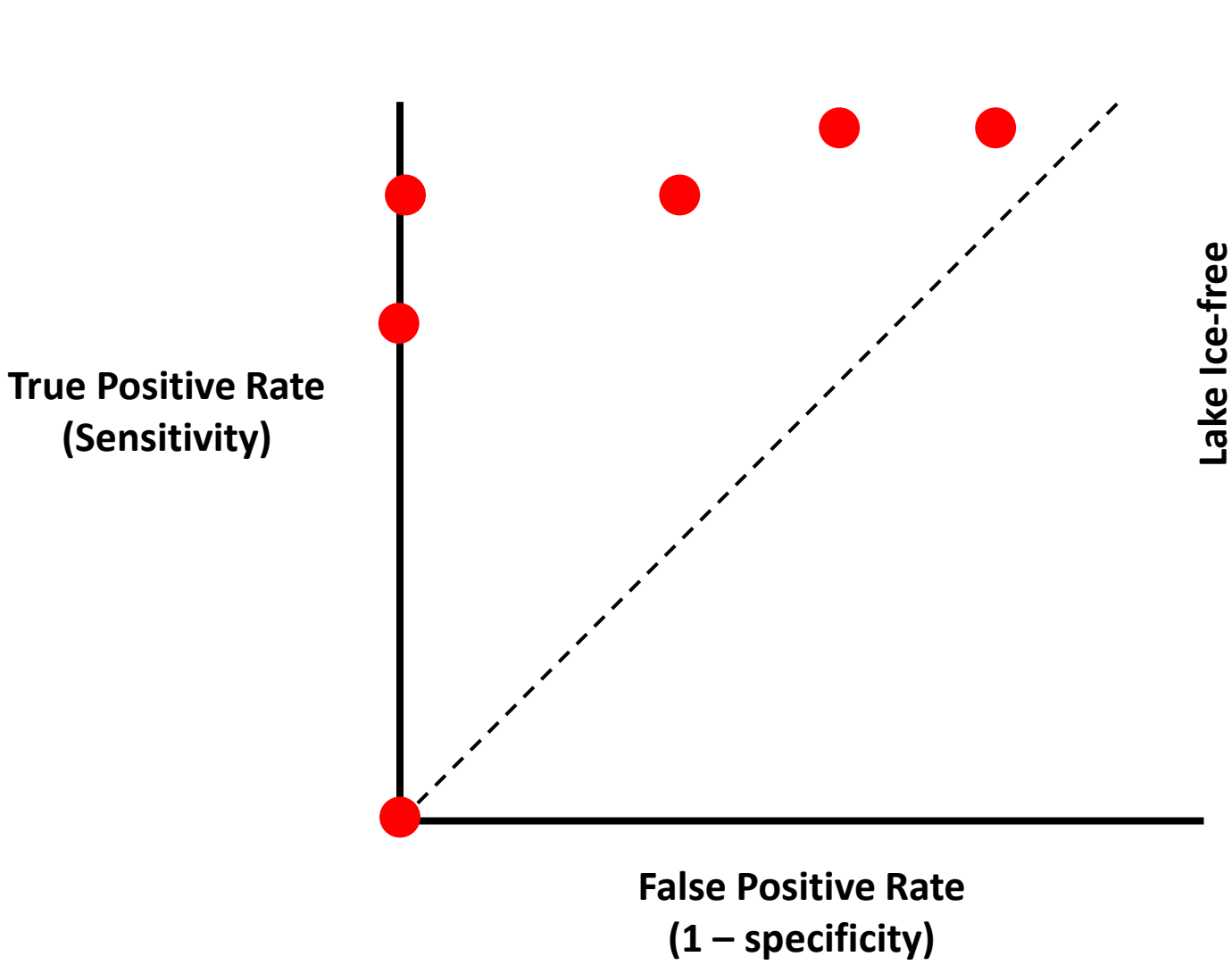



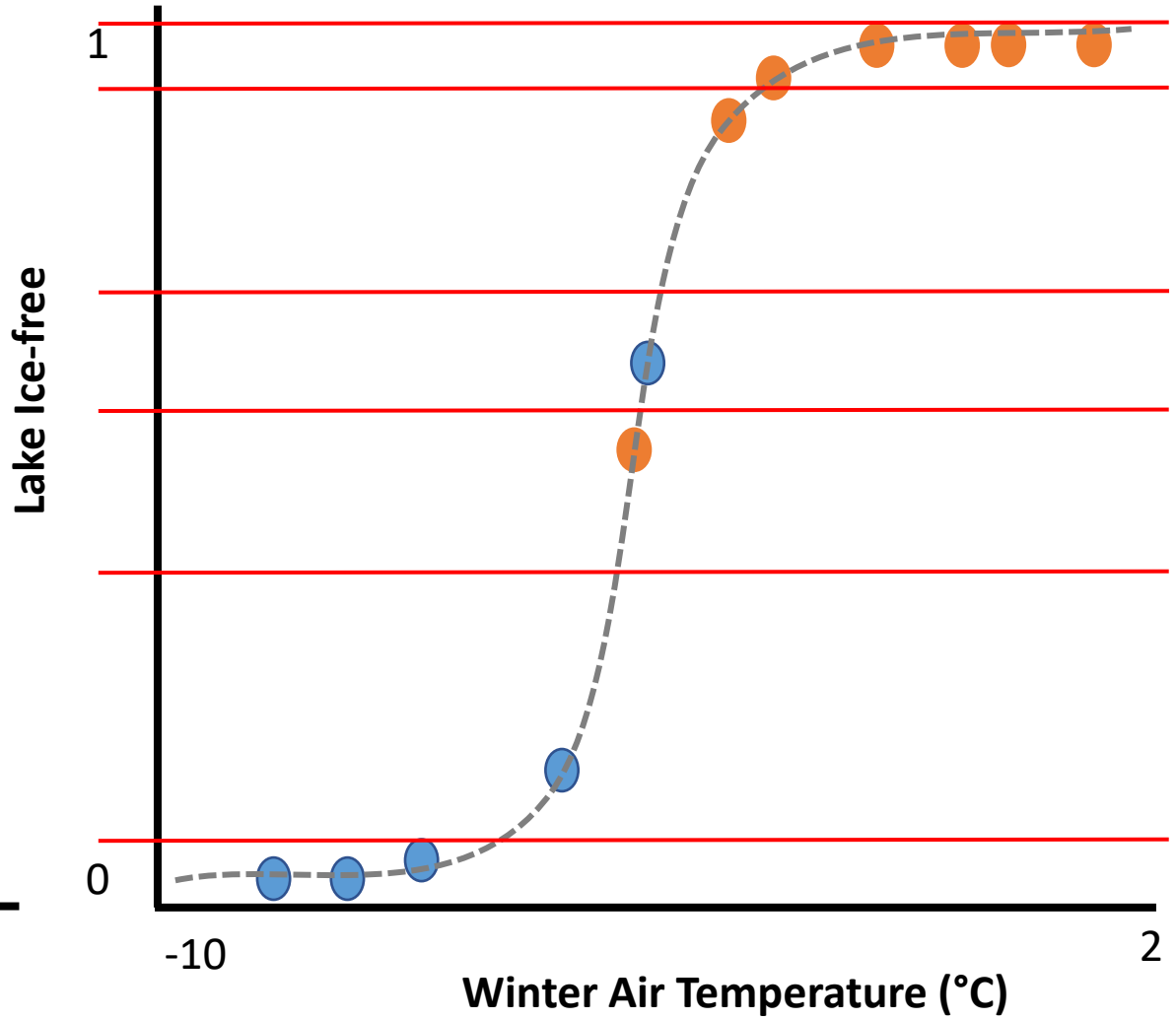
# Optimizing the threshold - ROC



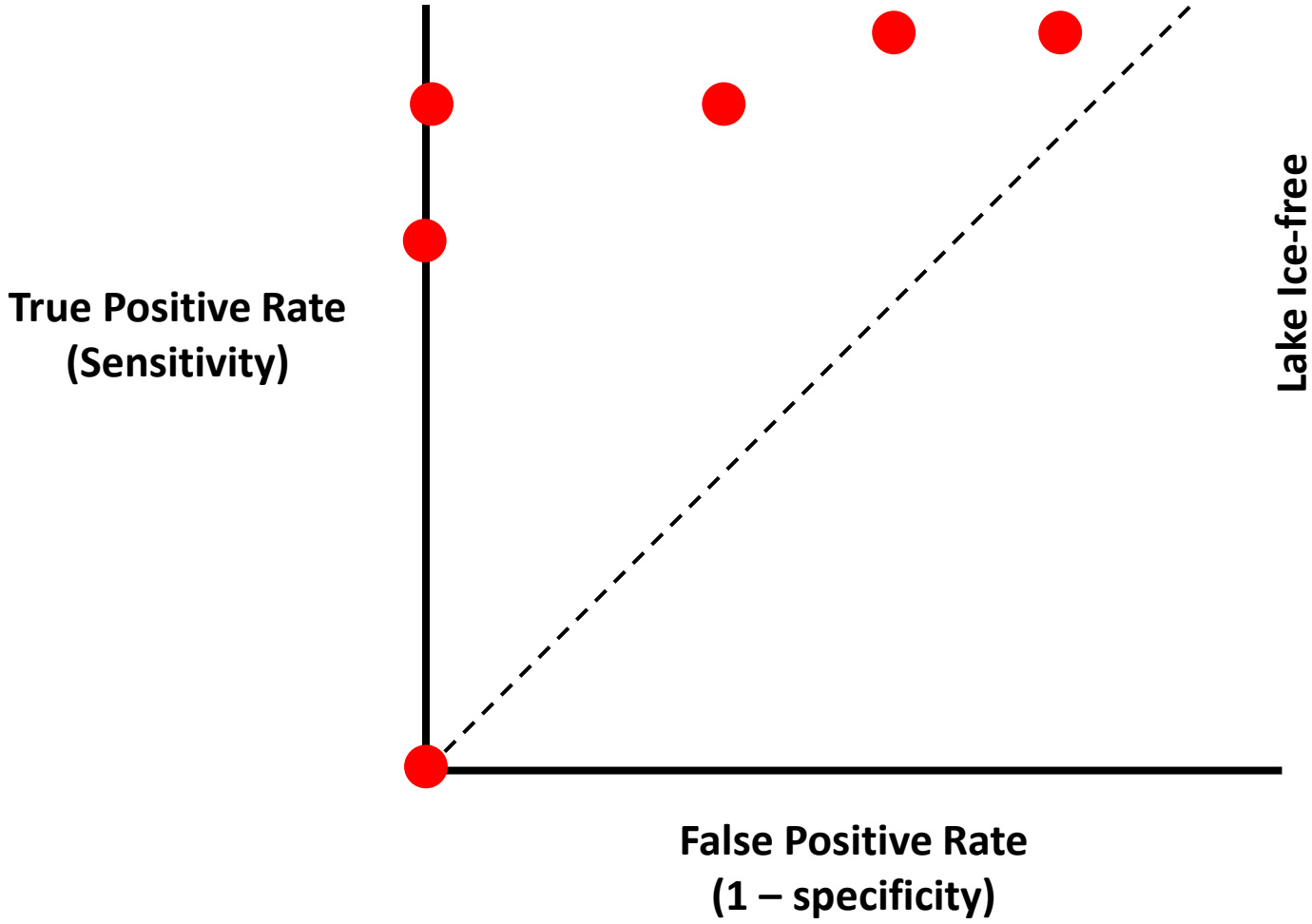
# Optimizing the threshold - ROC

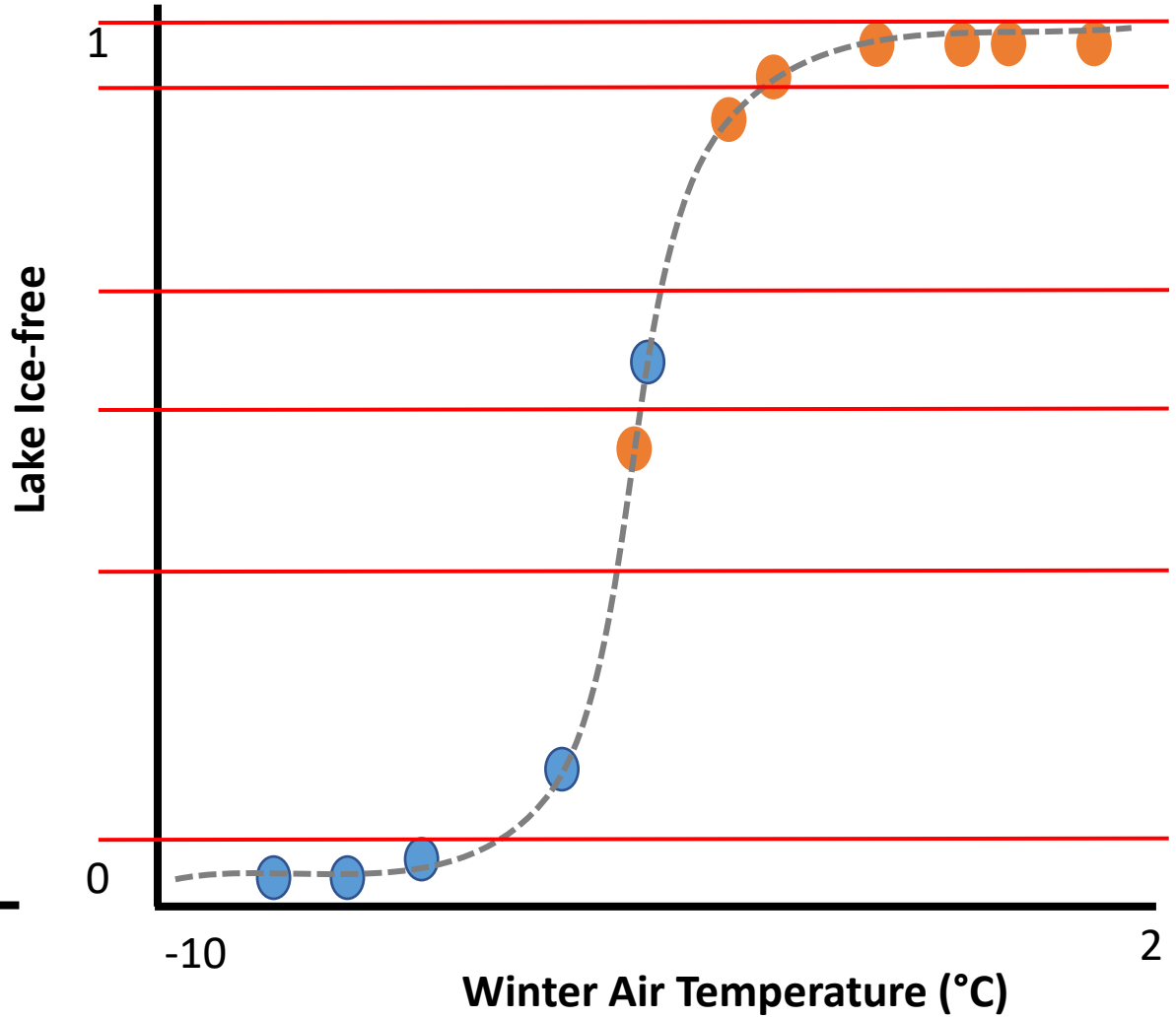


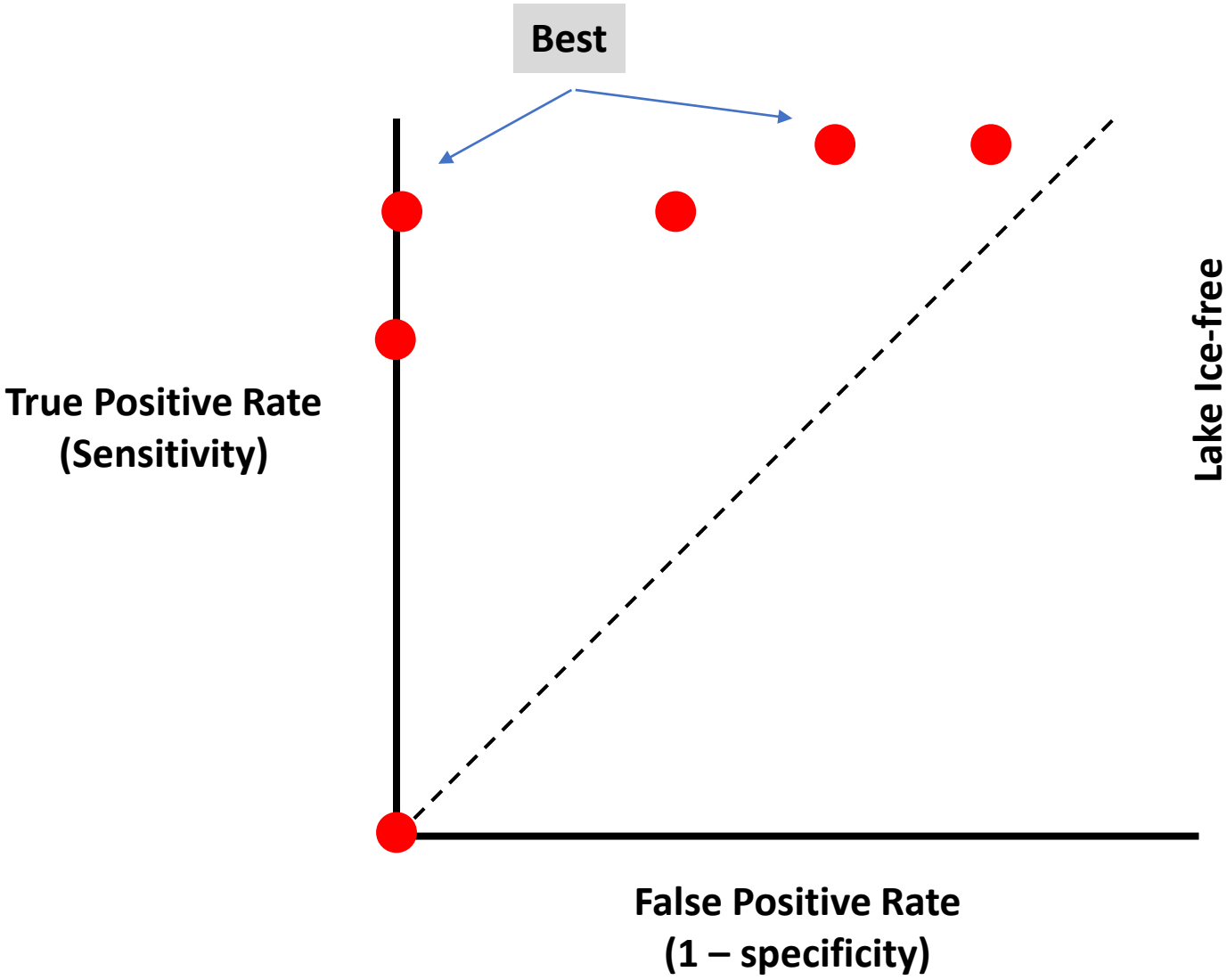
# Optimizing the threshold - ROC

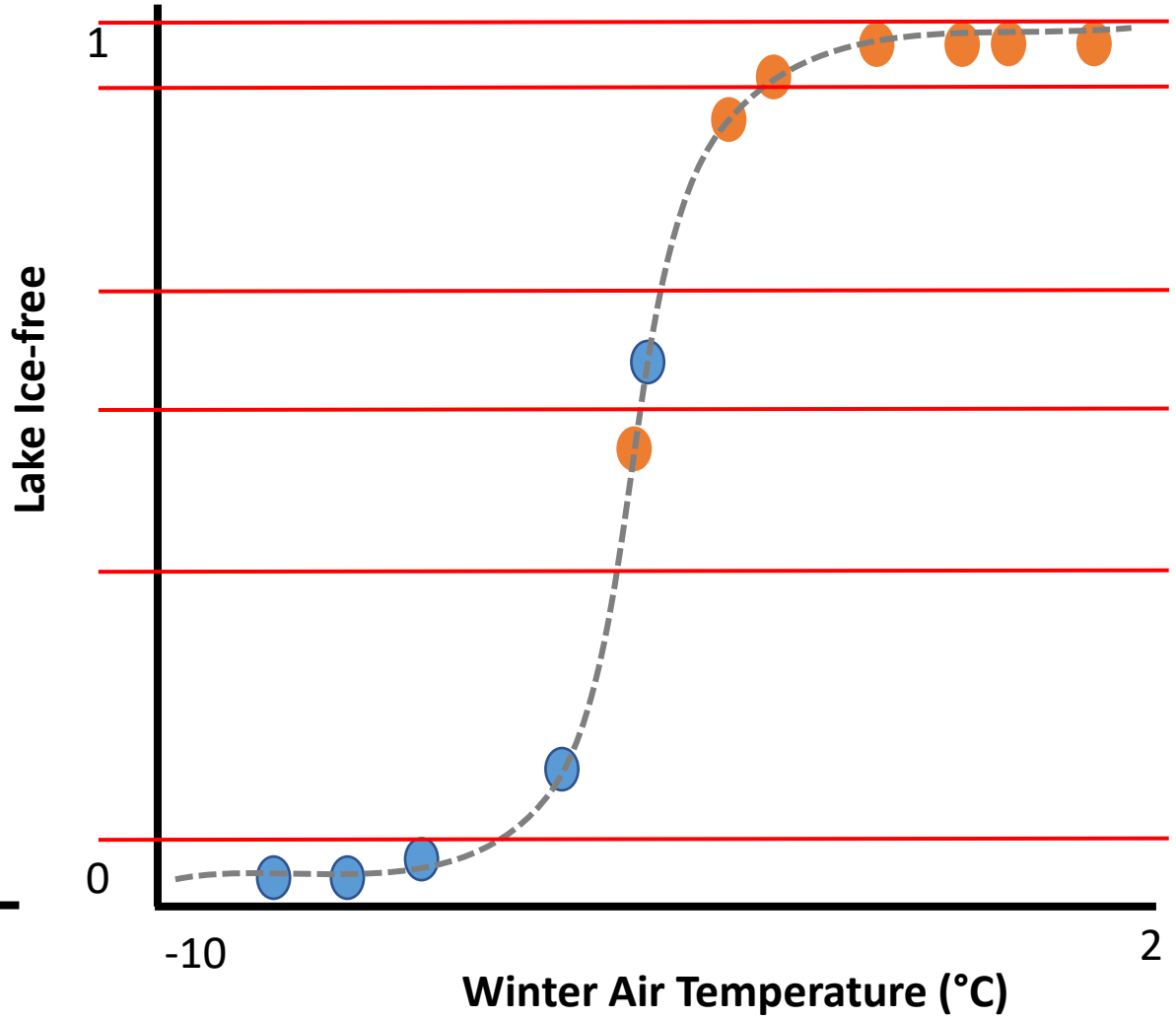
**Q: Which is the best?**



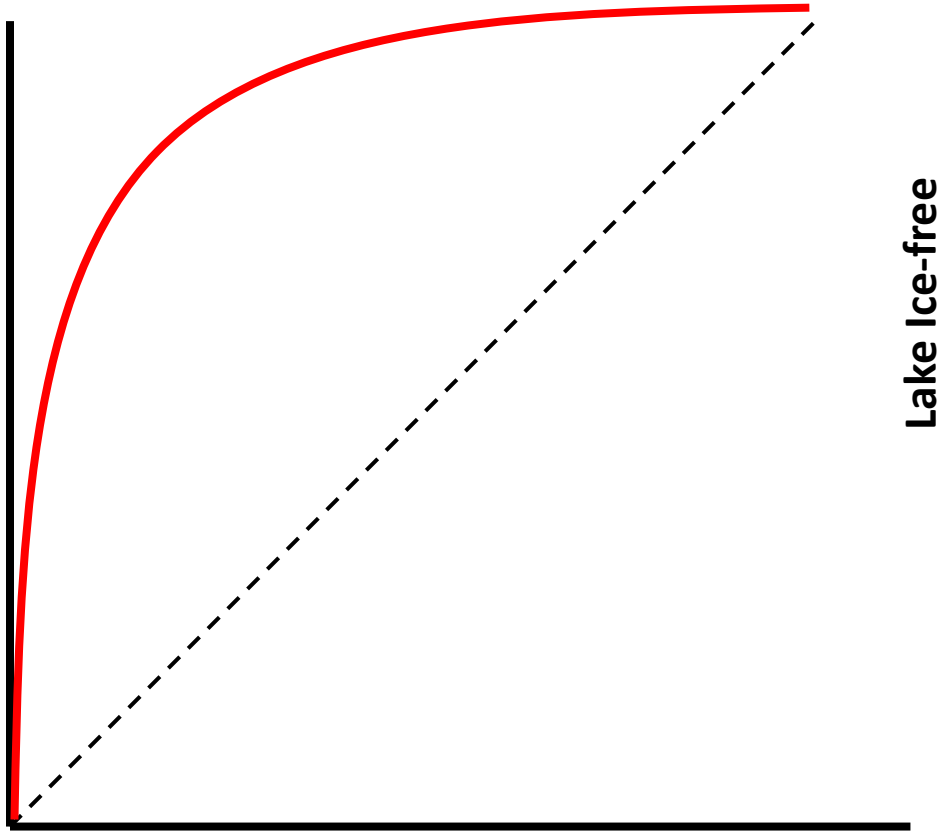



# Optimizing the threshold - ROC



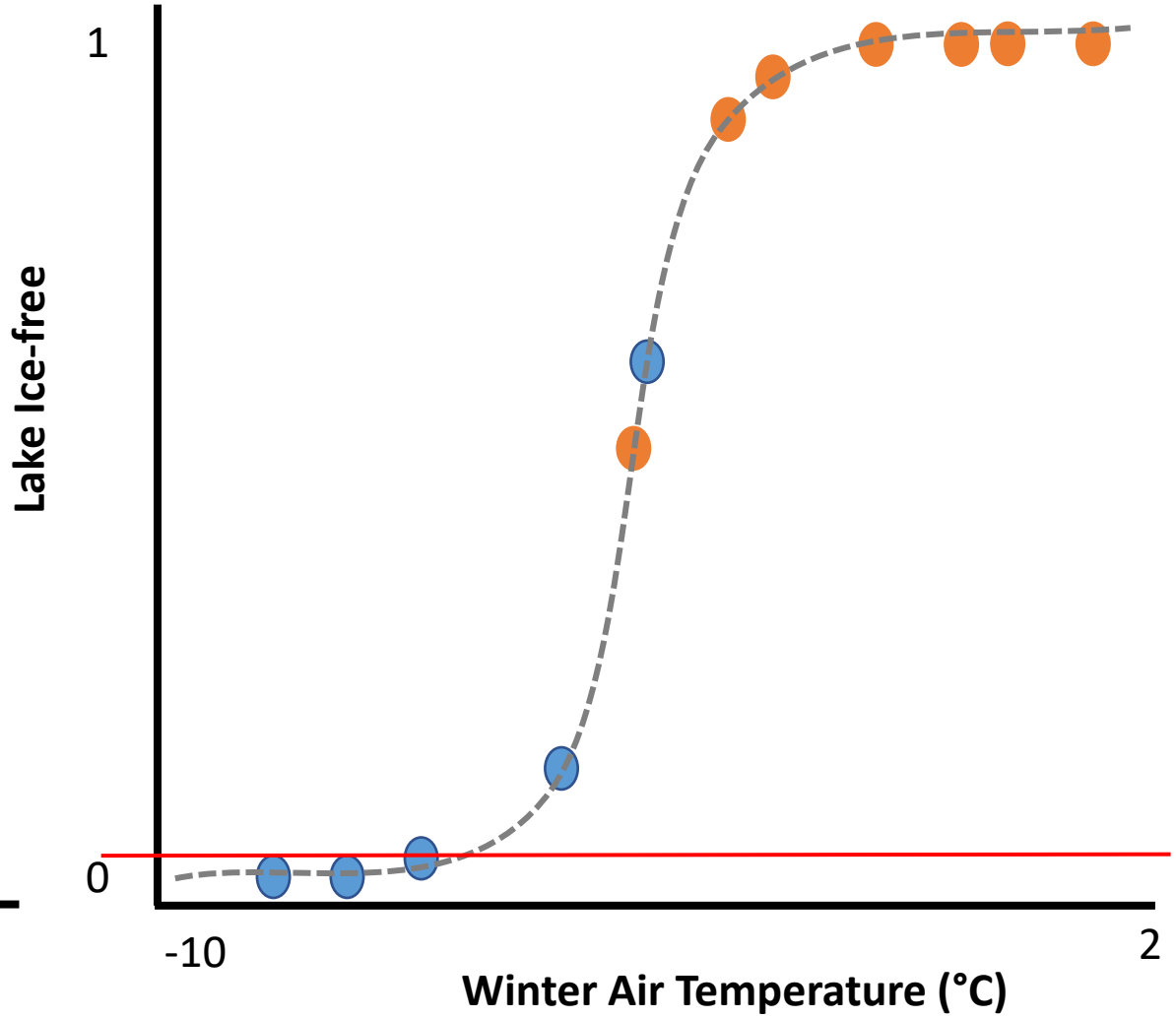



# Optimizing the threshold - ROC

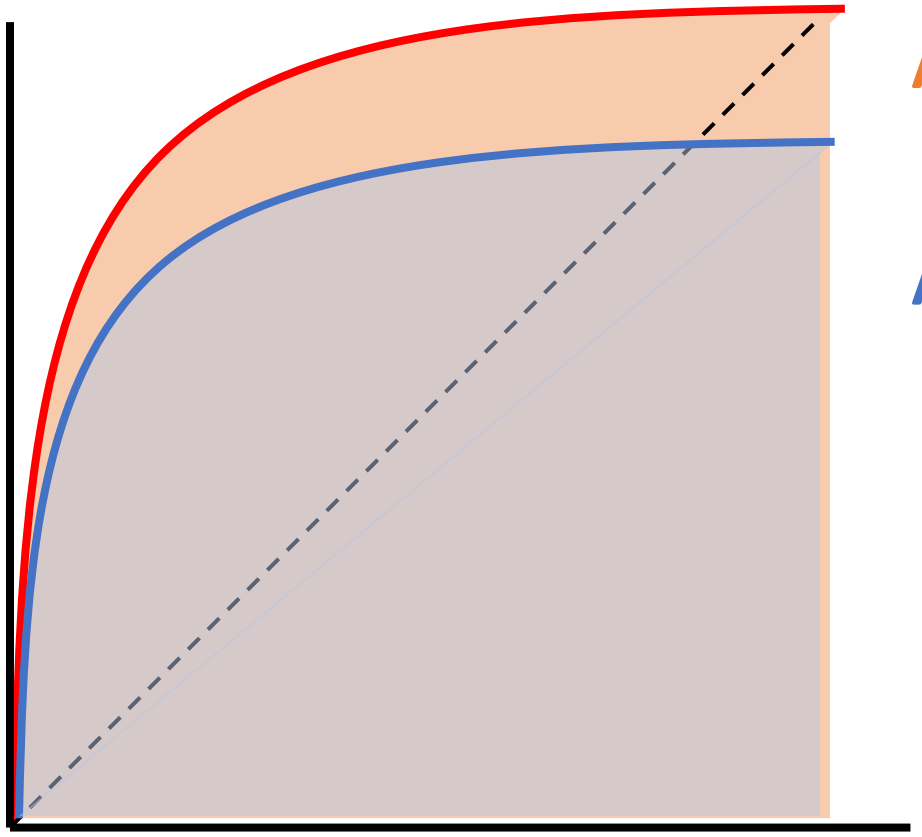


True Positive Rate
(Sensitivity)

False Positive Rate
(1 – specificity)



# Comparing models - AUC



AUC = 0.85

AUC = 0.76

True Positive Rate
(Sensitivity)

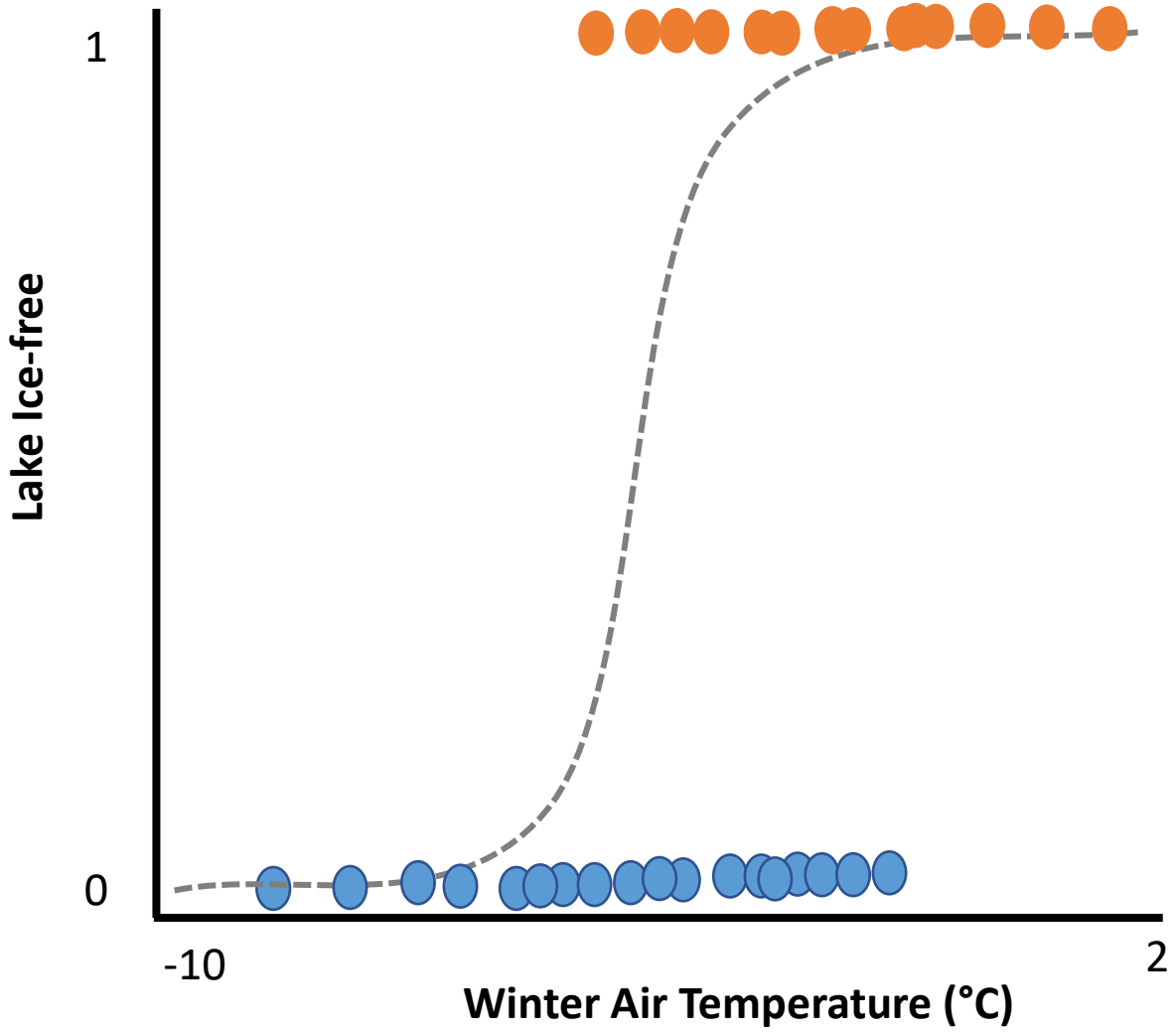
False Positive Rate
(1 – specificity)

# A note about overdispersion

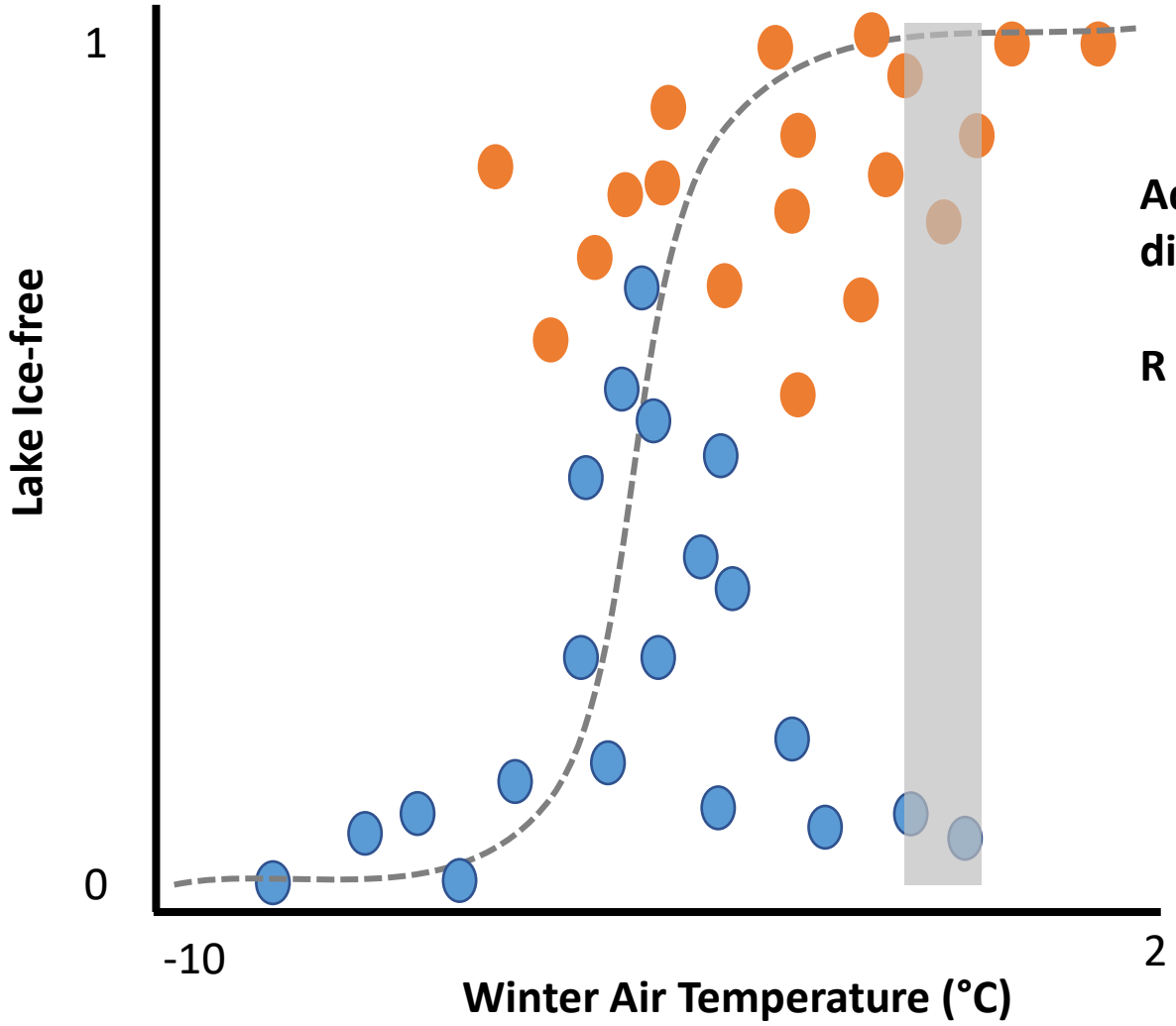
- Over dispersion when the variance exceeds the mean model fit

- Frequent in ecology data

# A note about overdispersion



# A note about overdispersion




**Adjust for this using a dispersion parameter**

**R - Quasi-binomial**

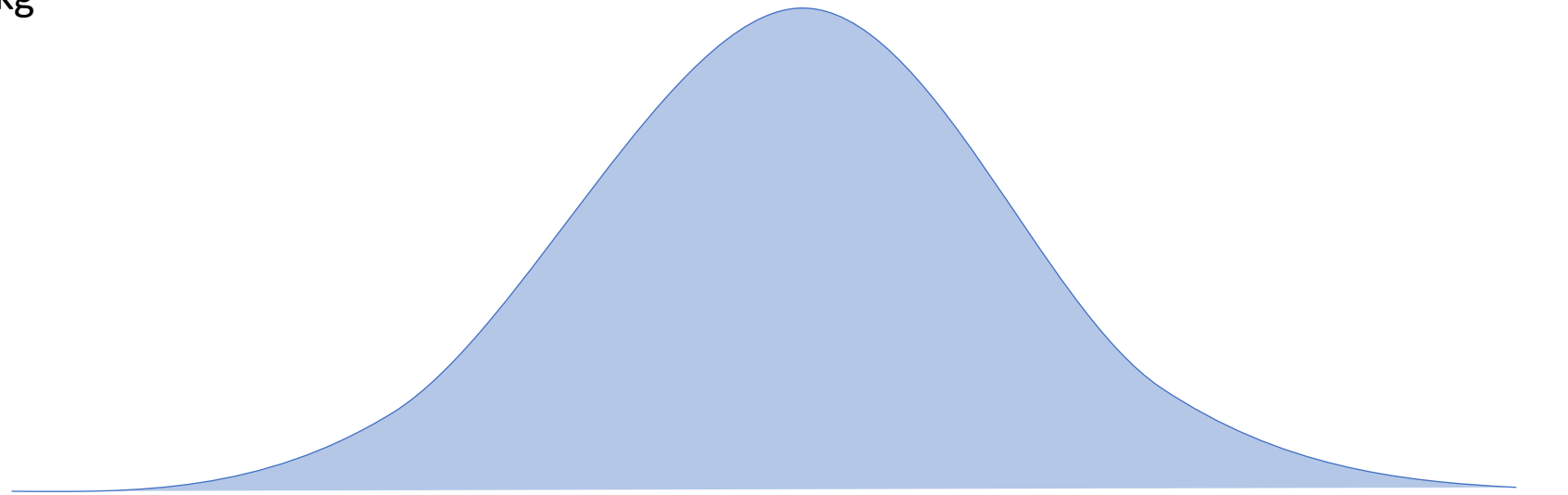
# How do logistic regressions relate to GLMs?

- 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

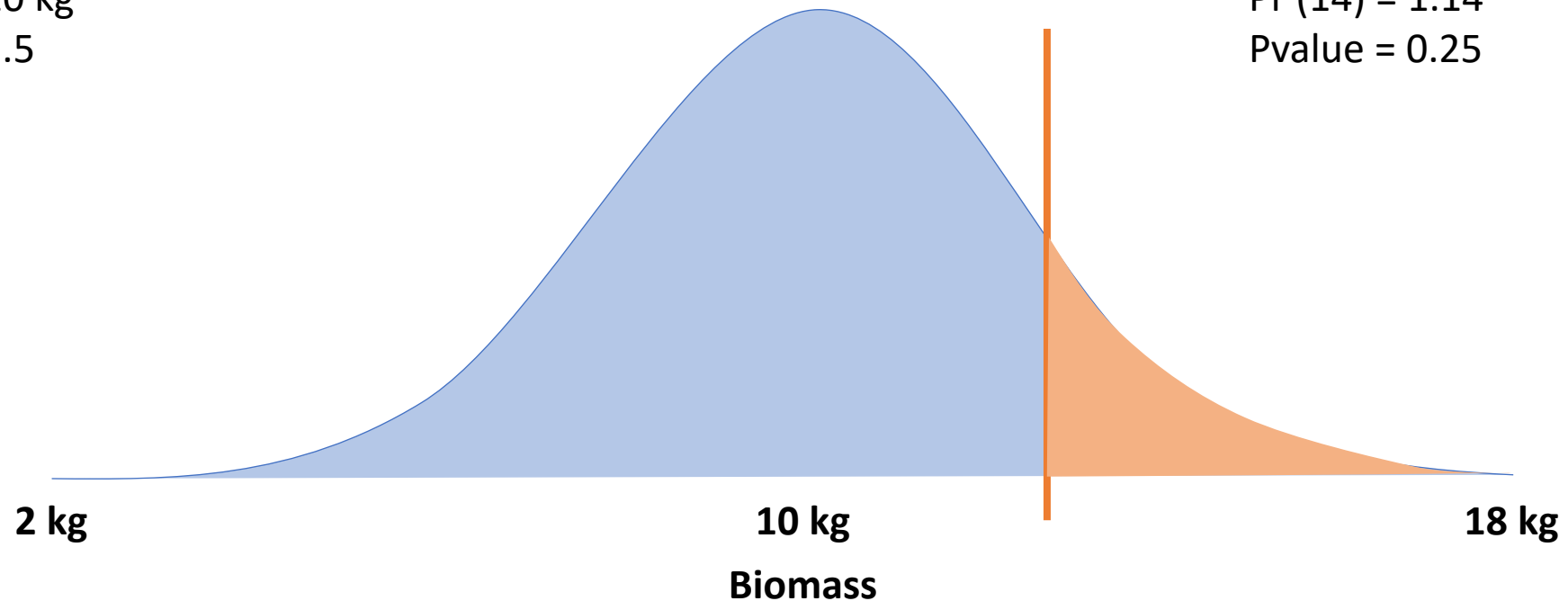


# GLM using a normal distribution

## **Probability** vs. Likelihood

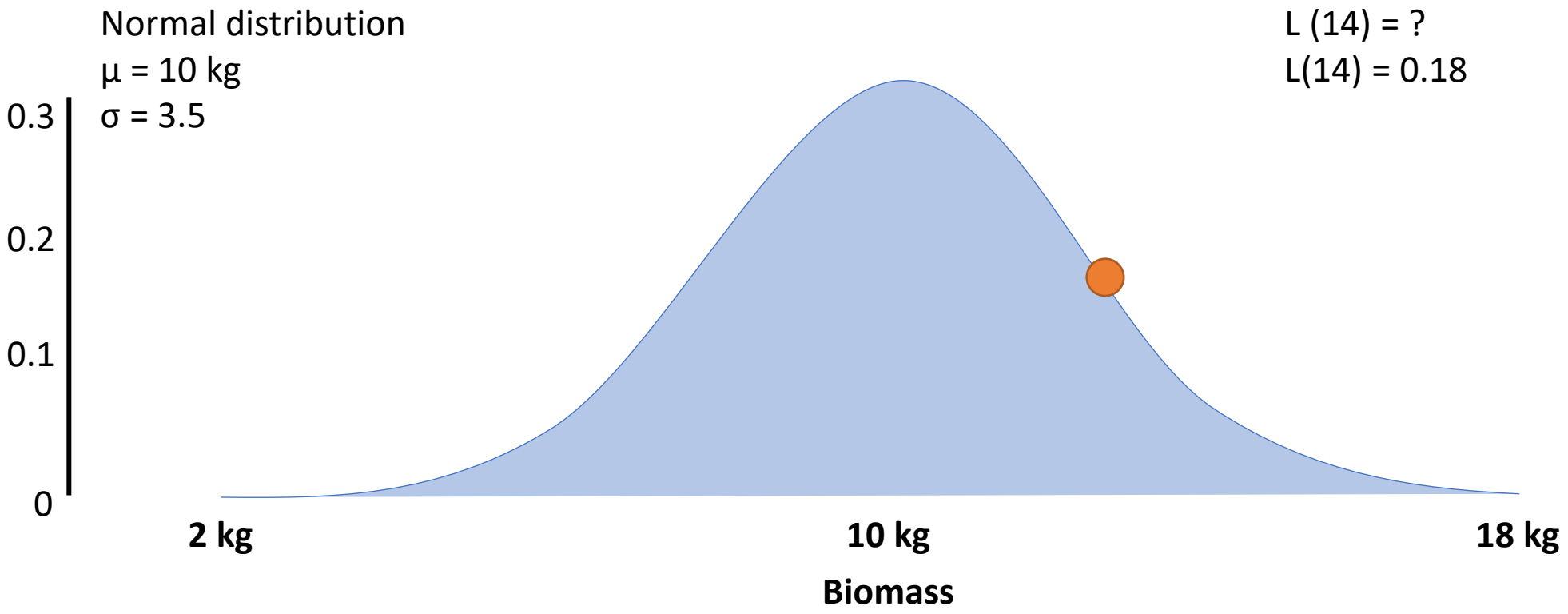
Normal distribution
μ = 10 kg
σ = 3.5

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



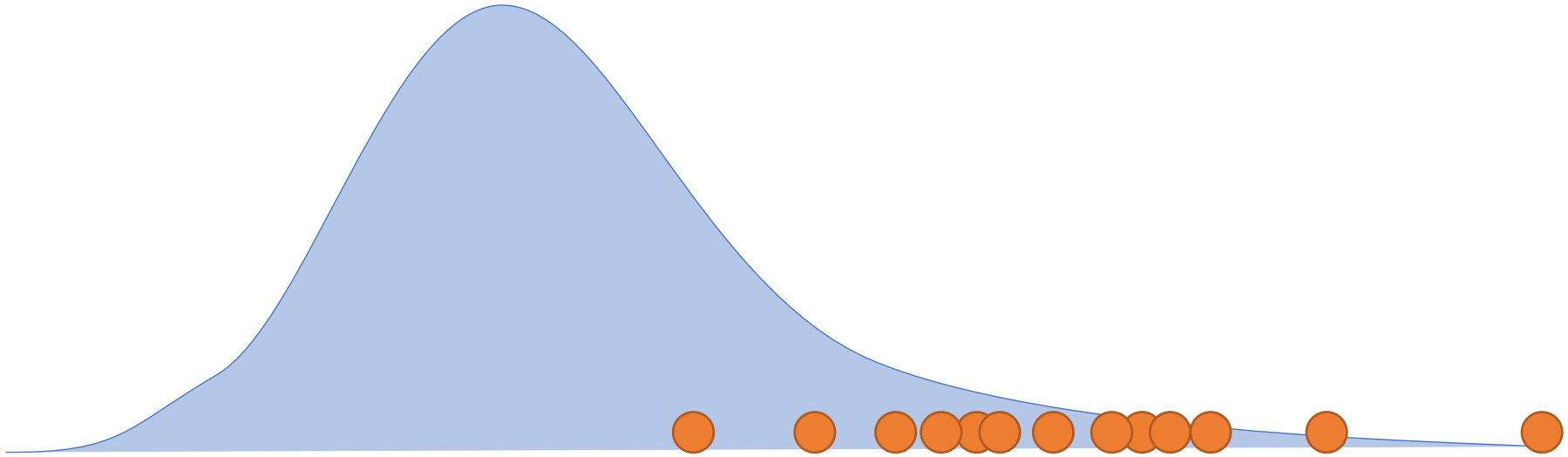
# GLM using a normal distribution

## Probability vs. **Likelihood**



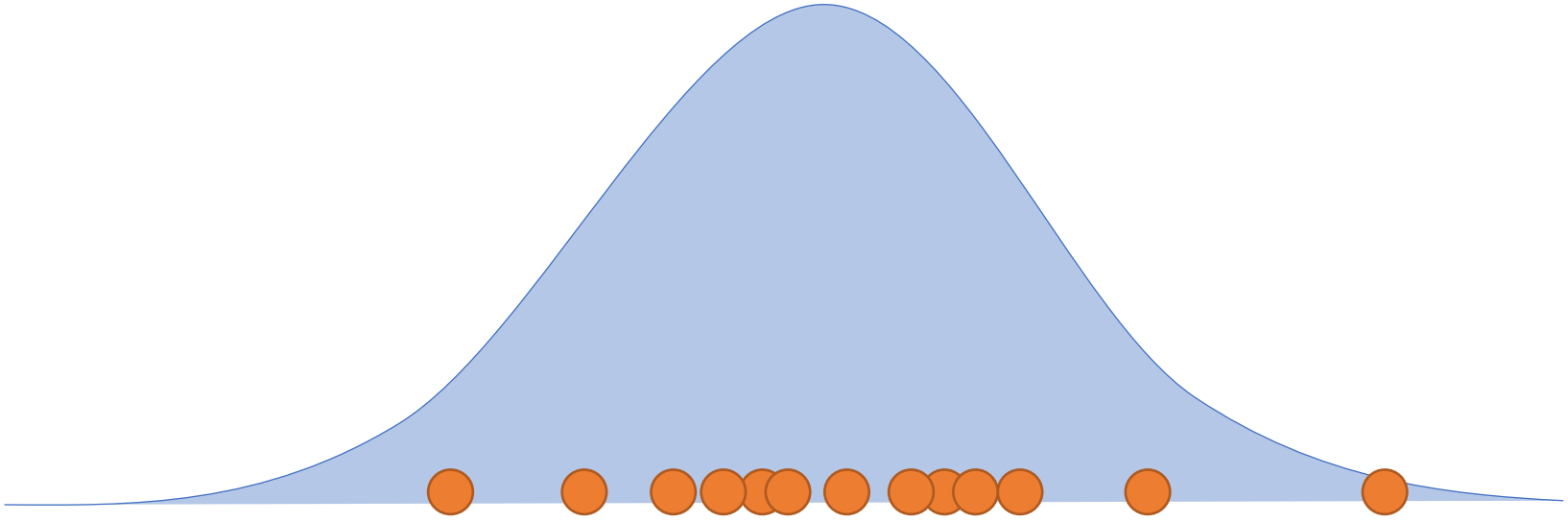
# Maximum likelihood for normal distribution

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



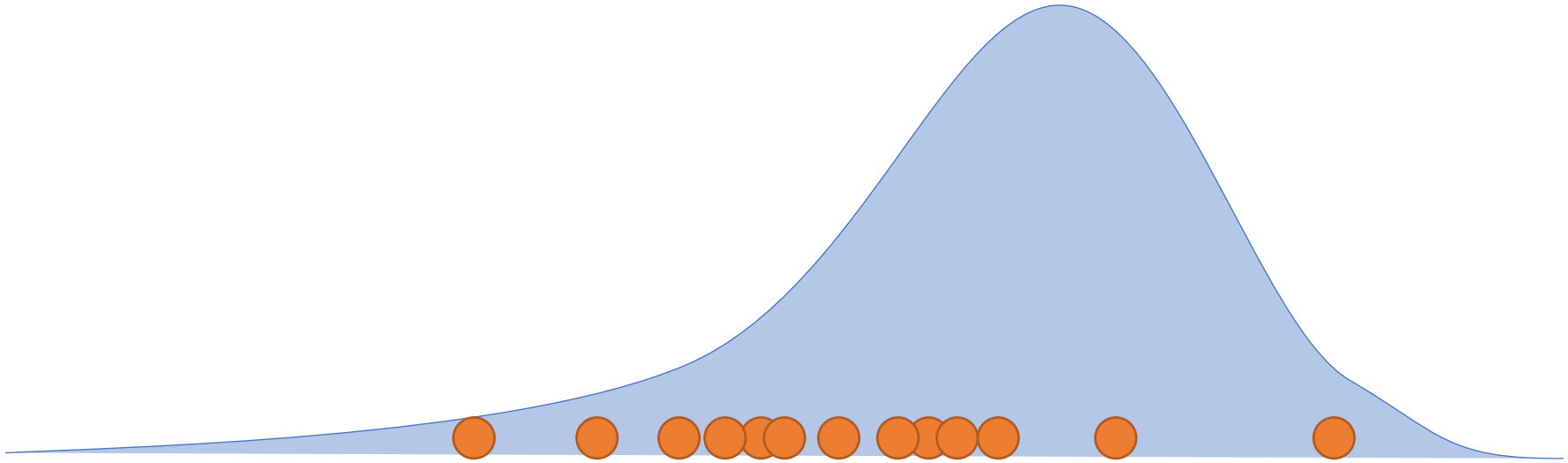
# Maximum likelihood for normal distribution

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

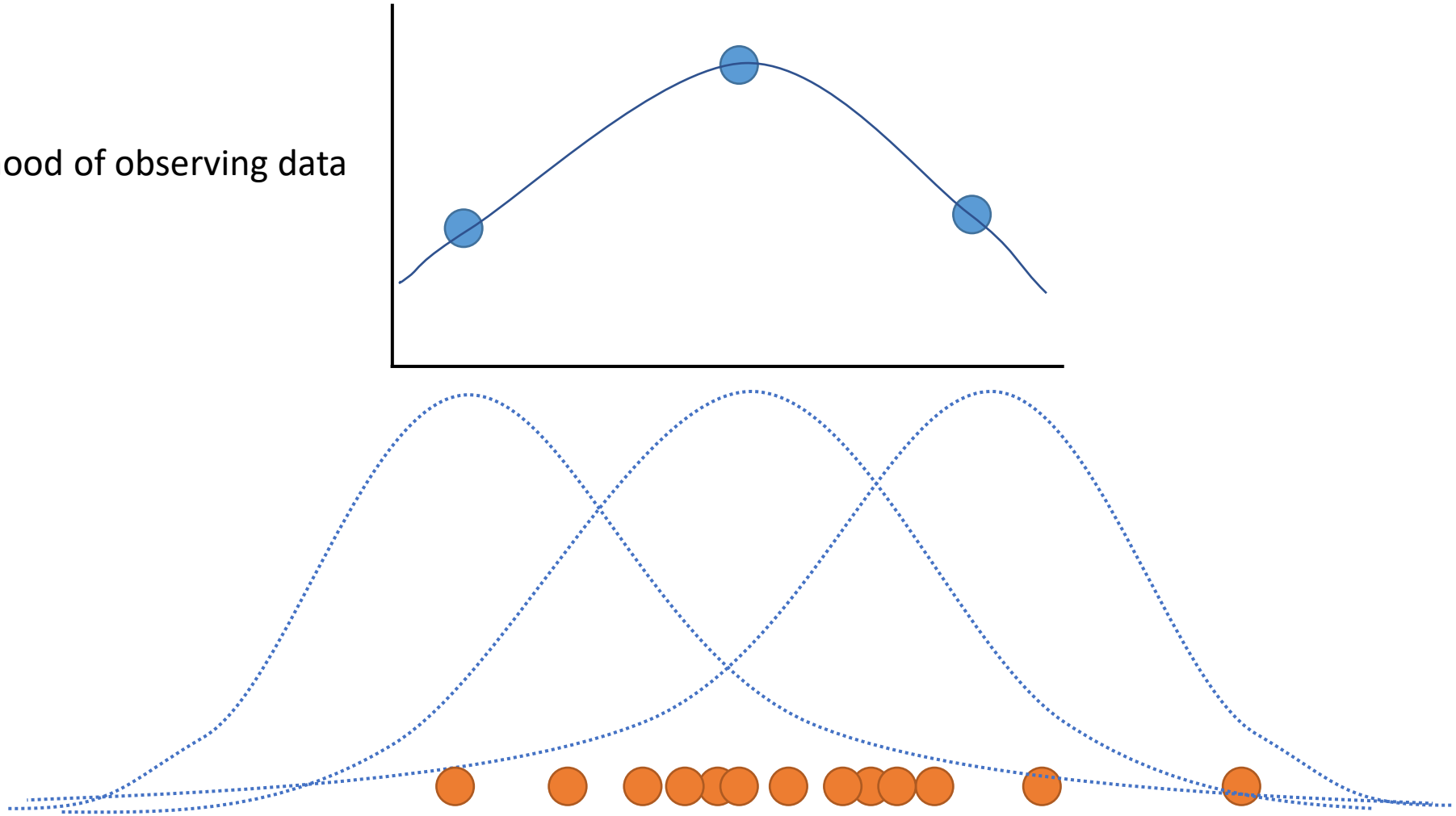


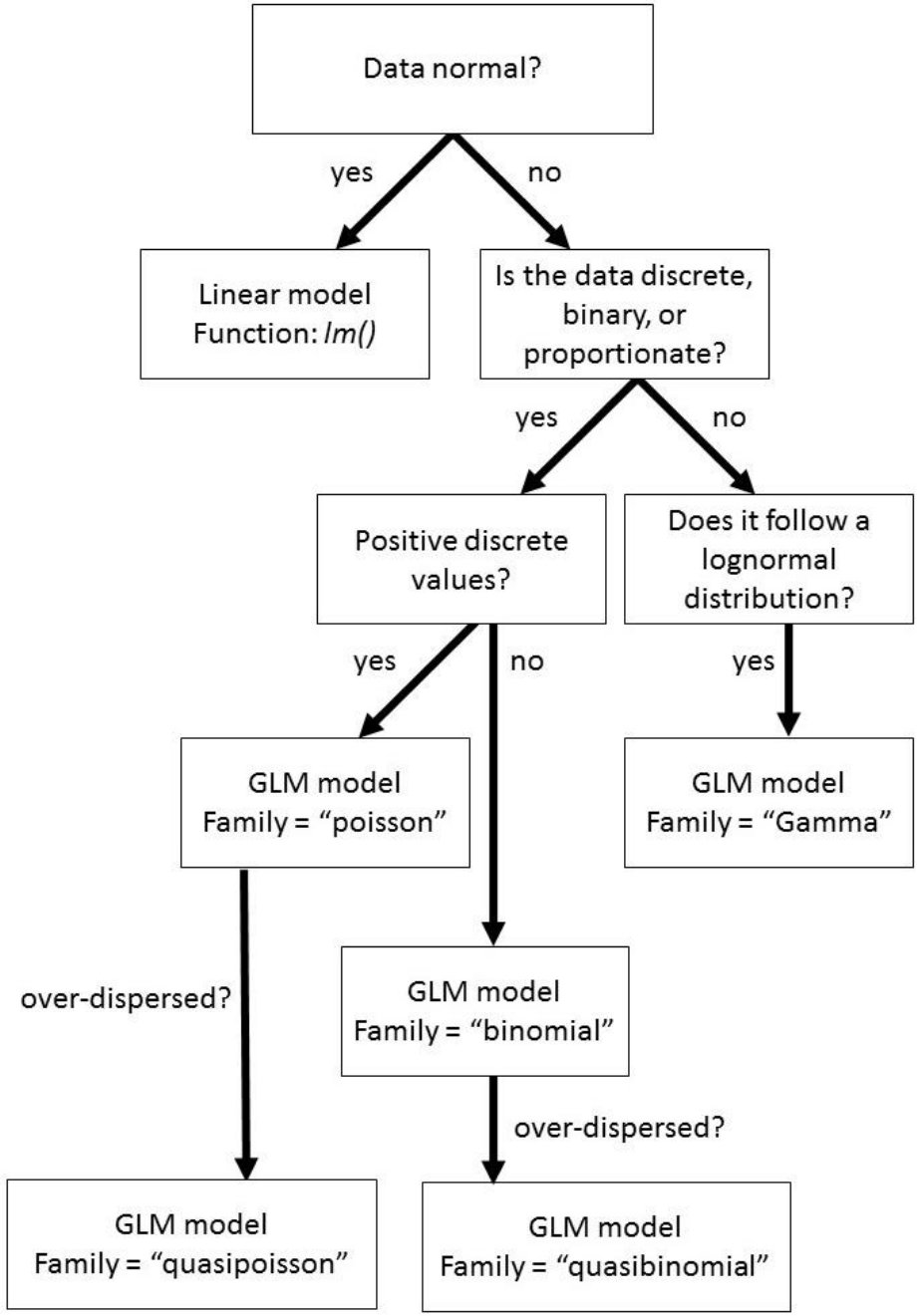
# Maximum likelihood for normal distribution

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



# Maximum likelihood for normal distribution

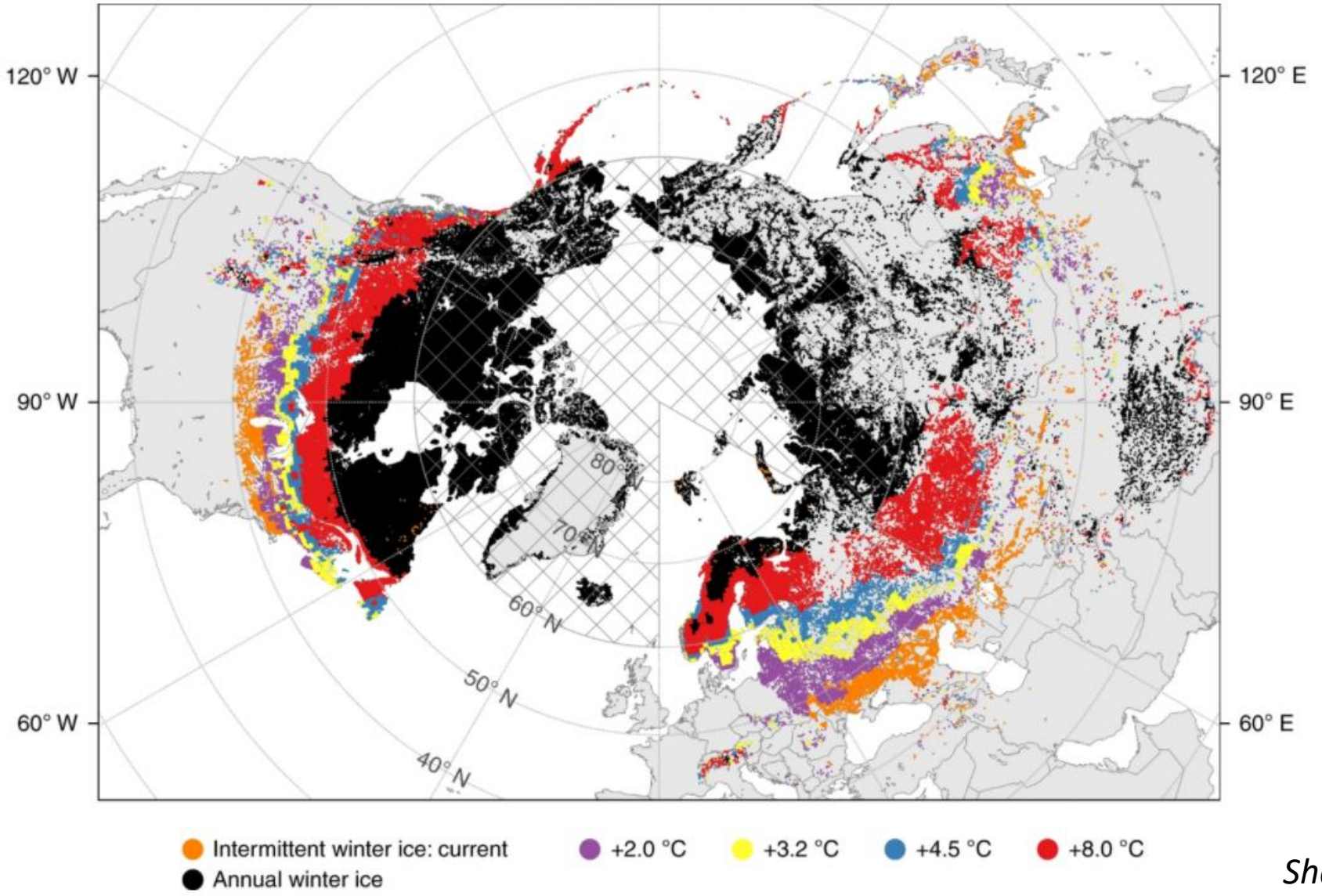





# A simple workflow for GLMs

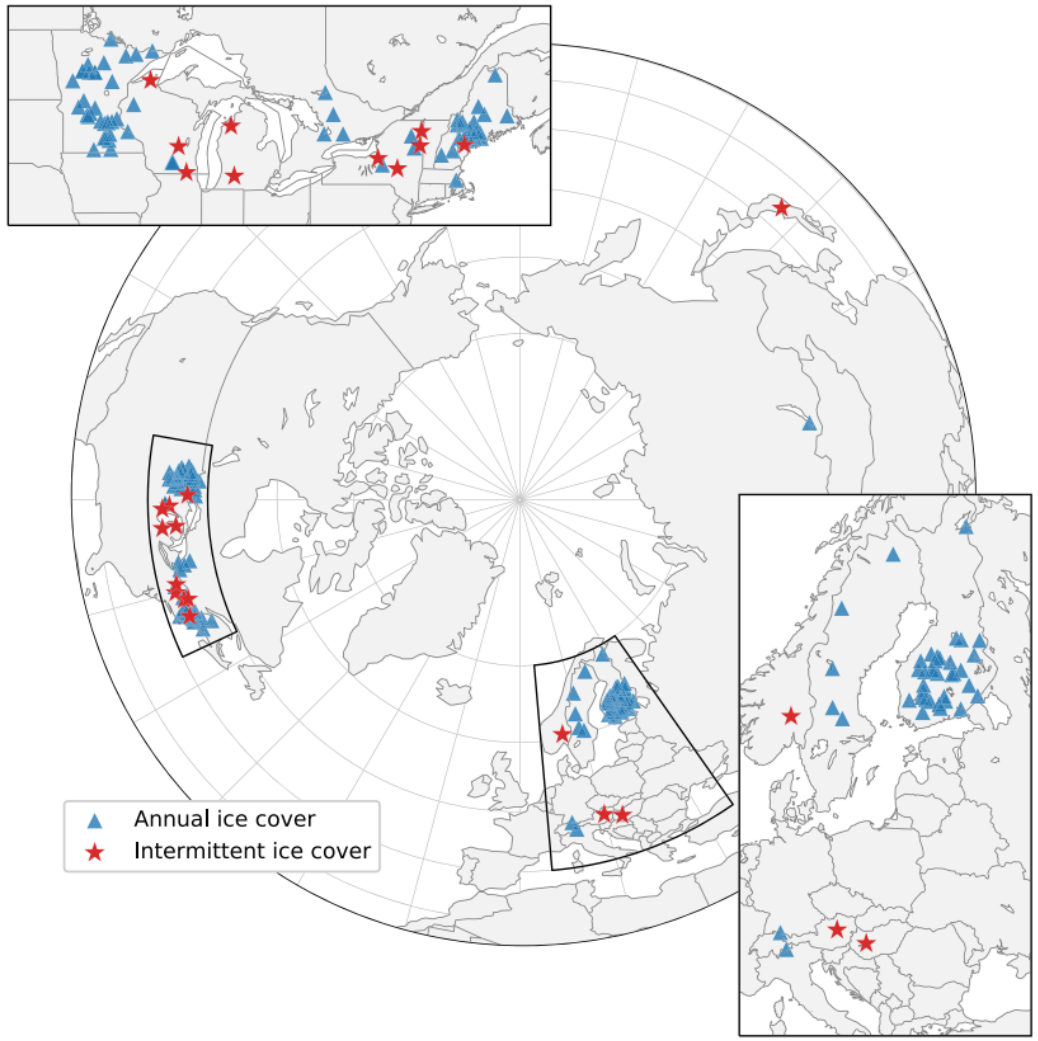
# Case Study: Logistic regression to identify extreme events

# Lake ice is threatened by climate change




*Sharma et al. 2019*

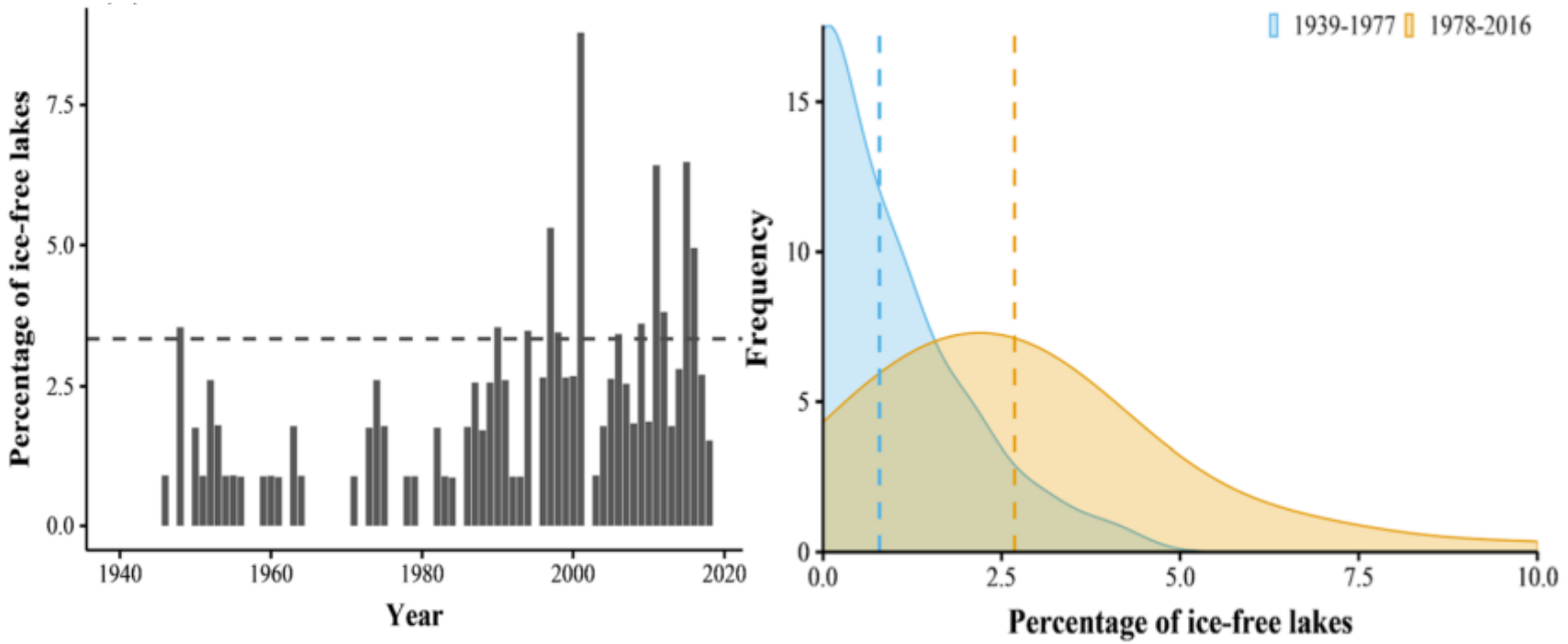
# Extreme events on lake ice




- 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

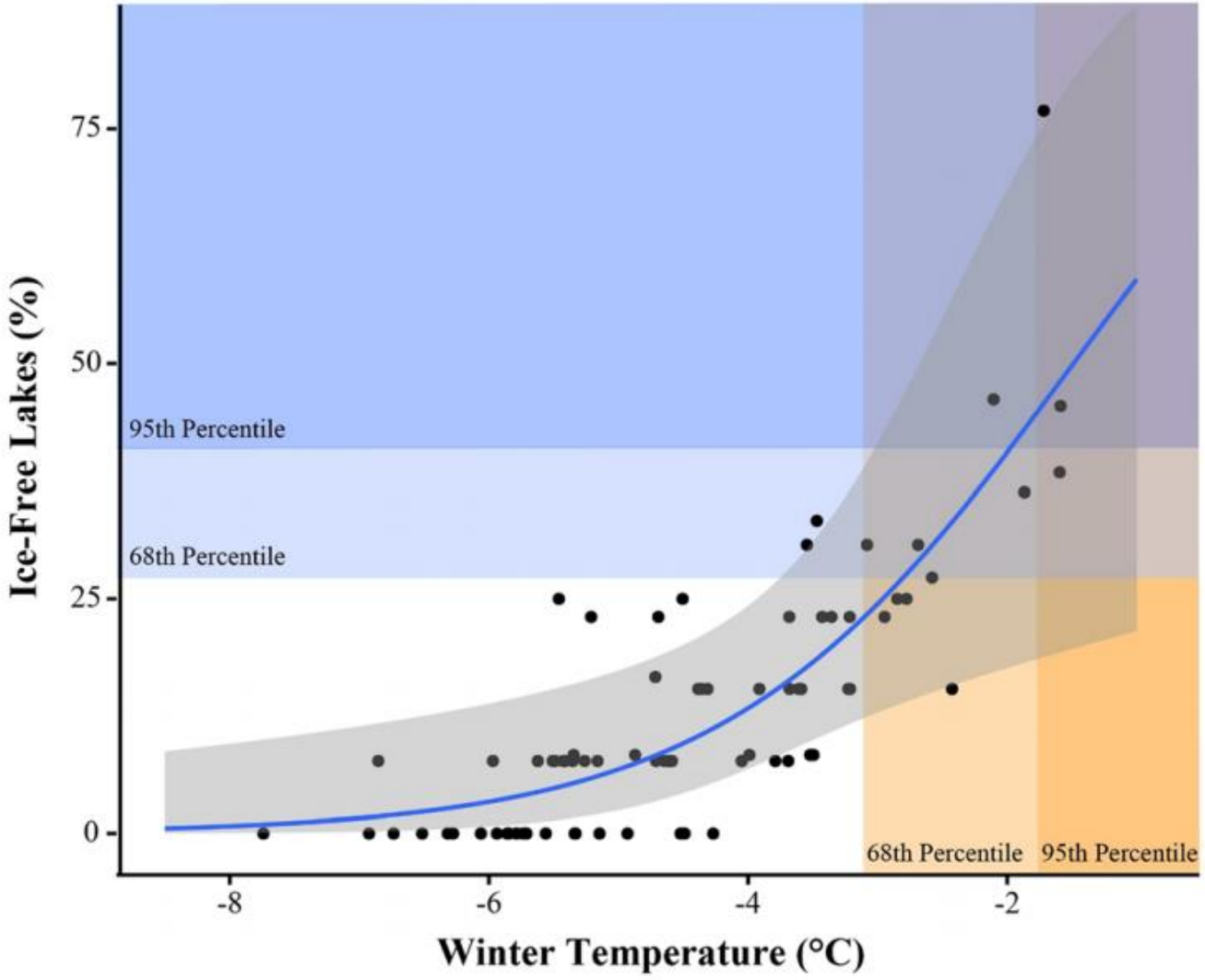



*Filazzola et al. 2020*

# Data of ice-free events

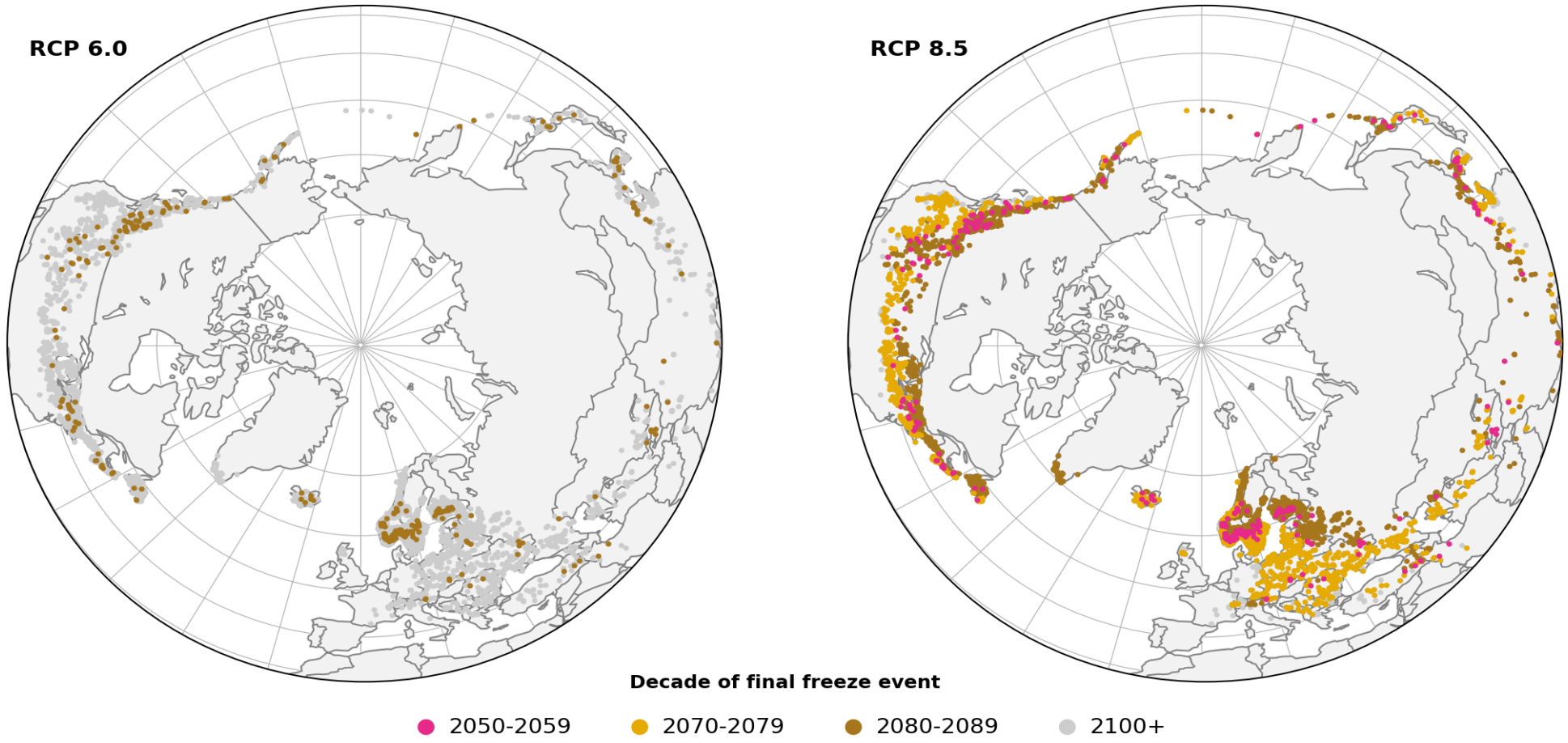
| Lake | Year | IceFree | Winter air temperature |
|---|---|---|---|
| Sebago Lake | 1995 | 1 | -1 |
| Sebago Lake | 1996 | 0 | -2.5 |
| Sebago Lake | 1997 | 0 | -2.9 |
| Sebago Lake | 1998 | 0 | -3.1 |
| George Lake | 1995 | 0 | -2.1 |
| George Lake | 1996 | 1 | -1.6 |
| George Lake | 1997 | 1 | -0.5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

# Extremes in temperature cause extremes in ice coverage




*Filazzola et al. 2020*

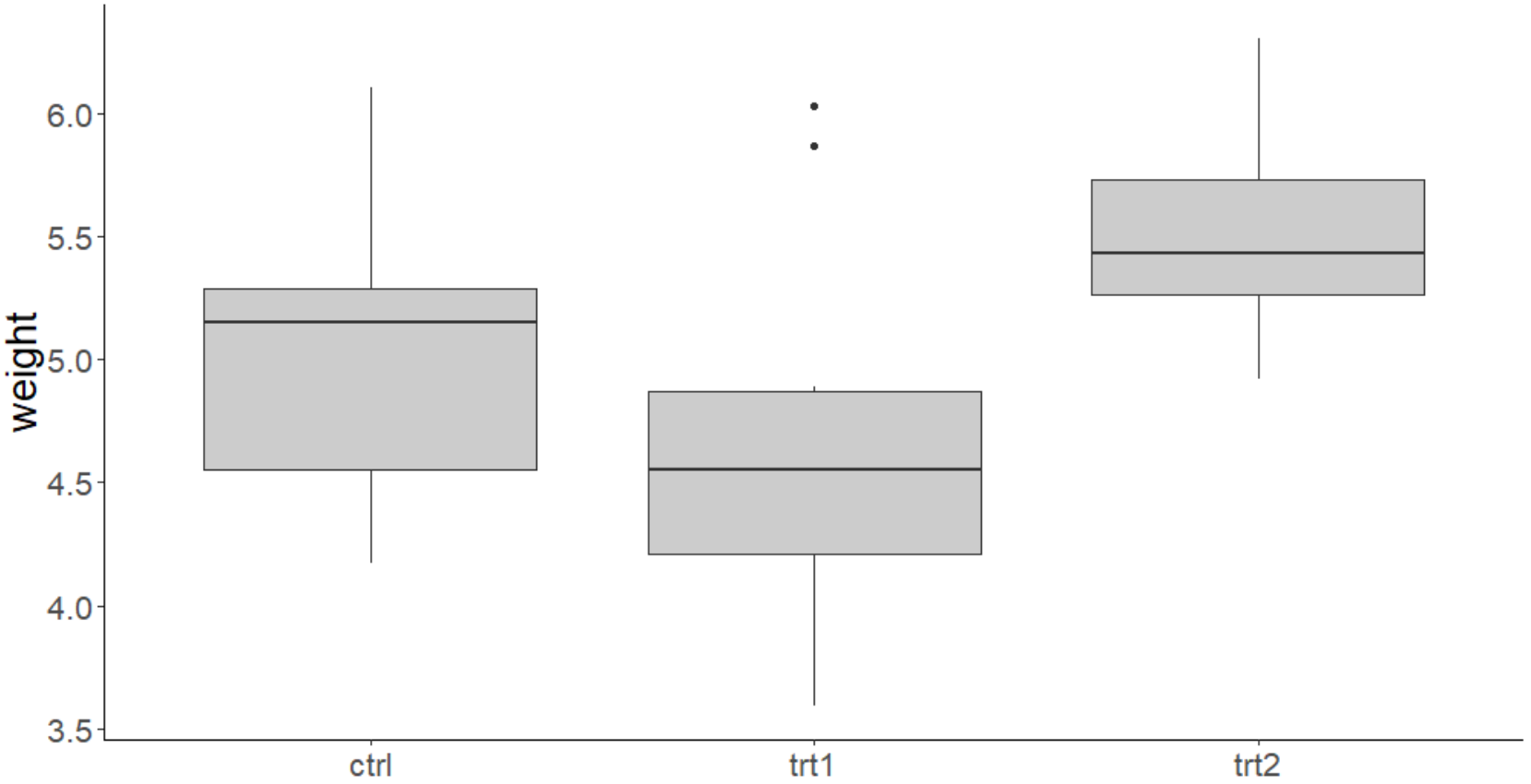
# The future loss of ice coverage




*Sharma et al. 2021*

Part C

# Post-hoc analyses




**GLM output**

$\chi^2 = 3.77$, df = 2, p = 0.007

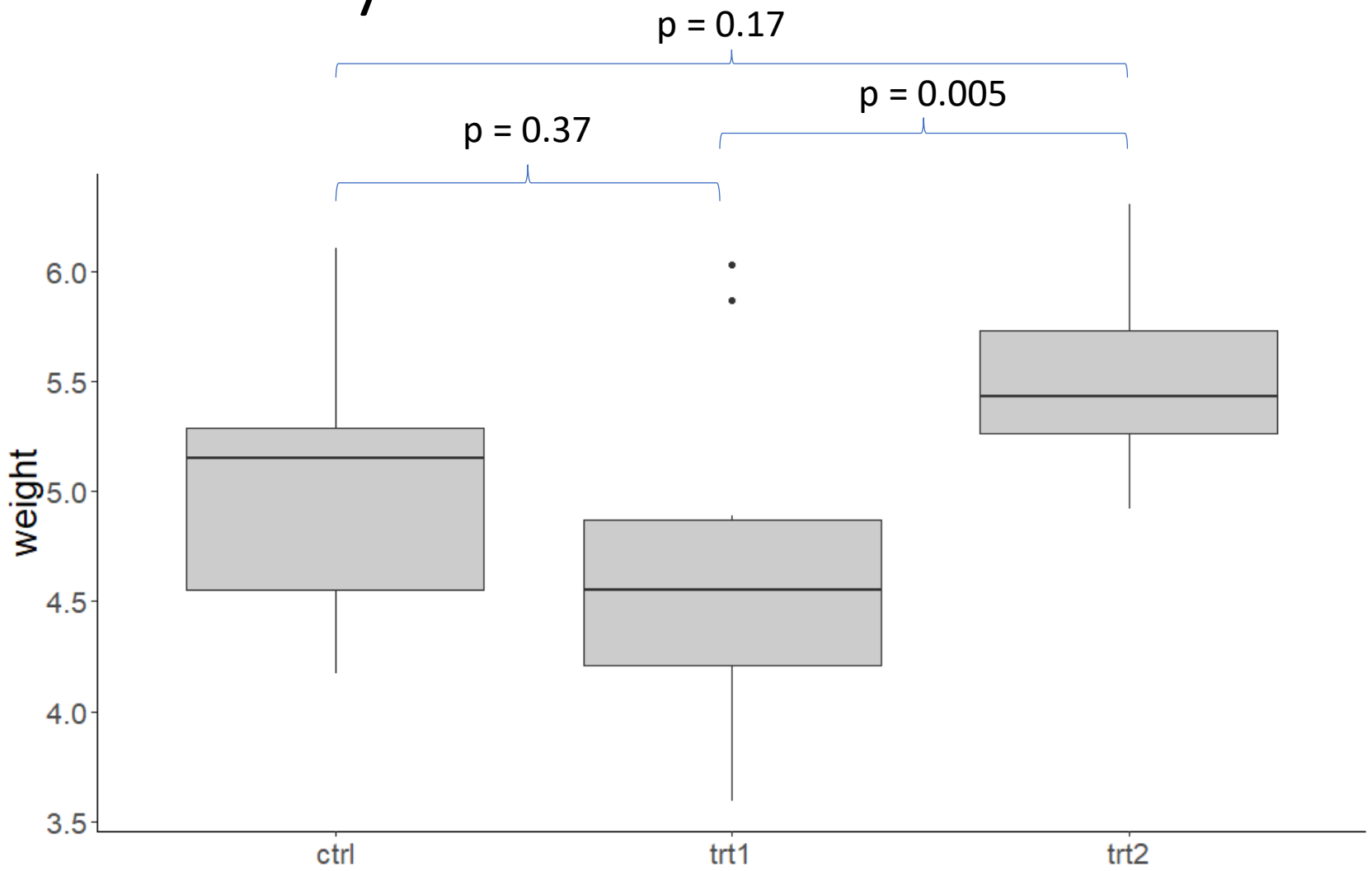
**Which groups are different?**

A – B

B – C

A – C

# Post-hoc analyses



# Generalized linear mixed models

- A GLM that includes both fixed effects and random effects
- **Fixed effects**: predictor to test for significance; constant among observations; estimated using ML in GLMs
- **Random effects:** predictor to account for variation among groups; estimates vary between observations; typically estimated based on a near-normal distribution
- All linear models have fixed effects, but not all have random effects.

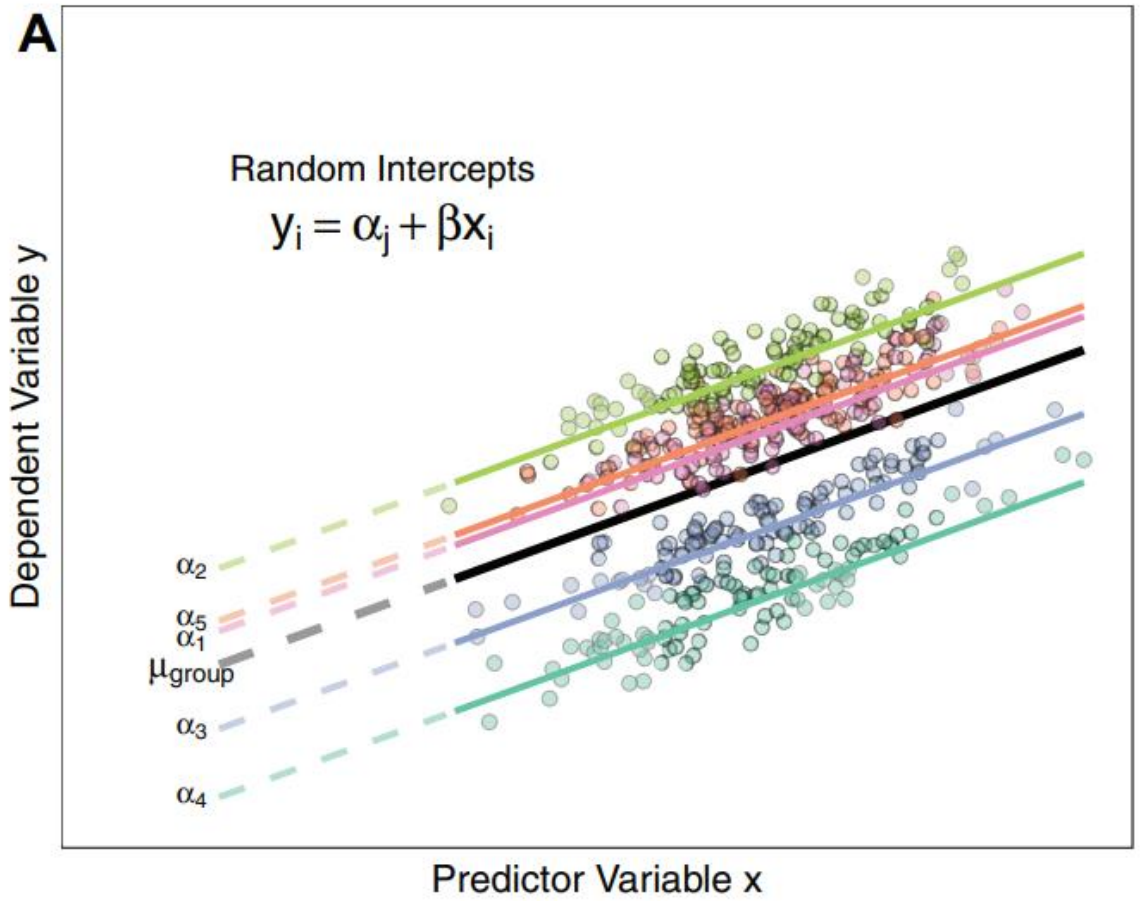
# When to use mixed models

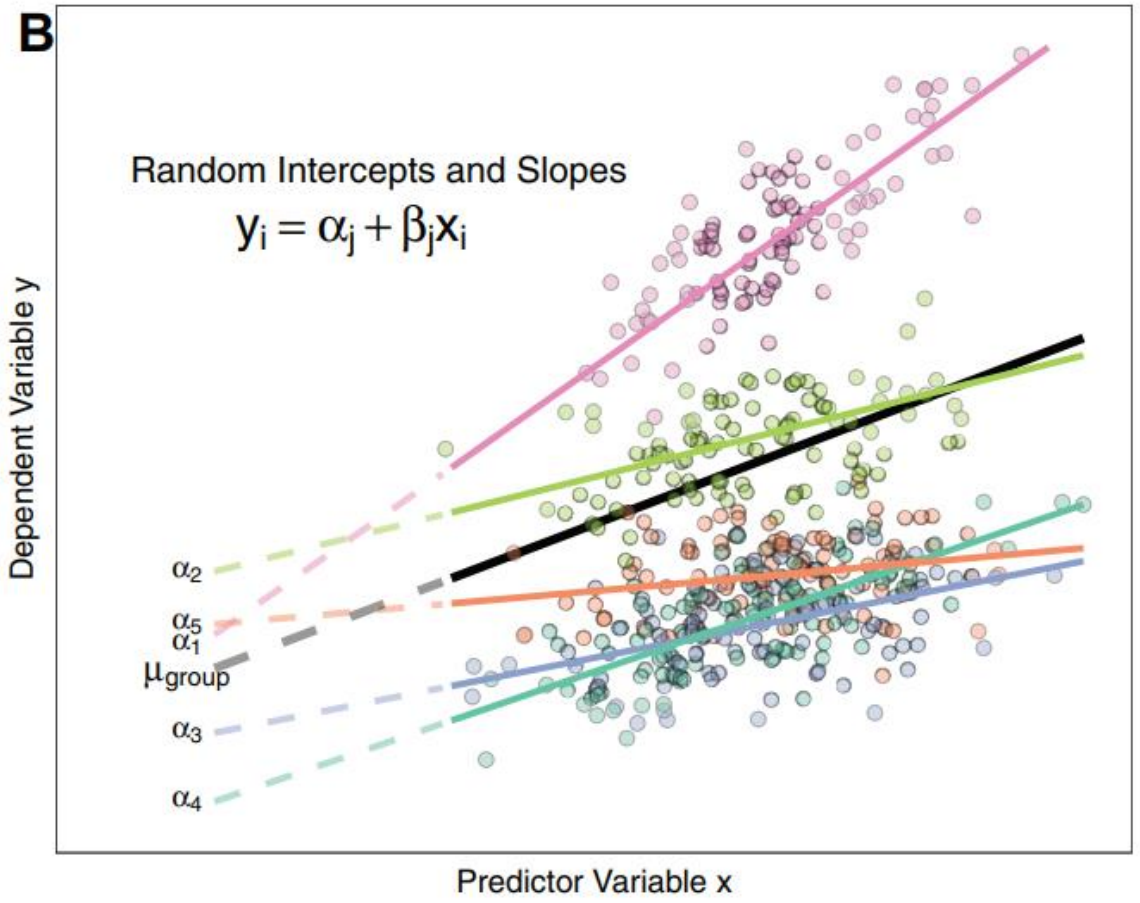
When trying to test for a particular effect while controlling for another variable

Examples:

- Testing the effect of a blood pressure drug on patients with different initial BP values
- Testing the effect of nutrient addition on plants surveyed in years with different growing conditions

# Types of mixed models




*Harrison et al., 2018*

# A great introduction to mixed models

PeerJ

## A brief introduction to mixed effects modelling and multi-model inference in ecology

Xavier A. Harrison[1], Lynda Donaldson[2,3], Maria Eugenia Correa-Cano[2], Julian Evans[4,5], David N. Fisher[4,6], Cecily E.D. Goodwin[2], Beth S. Robinson[2,7], David J. Hodgson[4] and Richard Inger[2,4]

[1] Institute of Zoology, Zoological Society of London, London, UK
[2] Environment and Sustainability Institute, University of Exeter, Penryn, UK
[3] Wildfowl and Wetlands Trust, Slimbridge, Gloucestershire, UK
[4] Centre for Ecology and Conservation, University of Exeter, Penryn, UK
[5] Department of Biology, University of Ottawa, Ottawa, ON, Canada
[6] Department of Integrative Biology, University of Guelph, Guelph, ON, Canada
[7] WildTeam Conservation, Padstow, UK

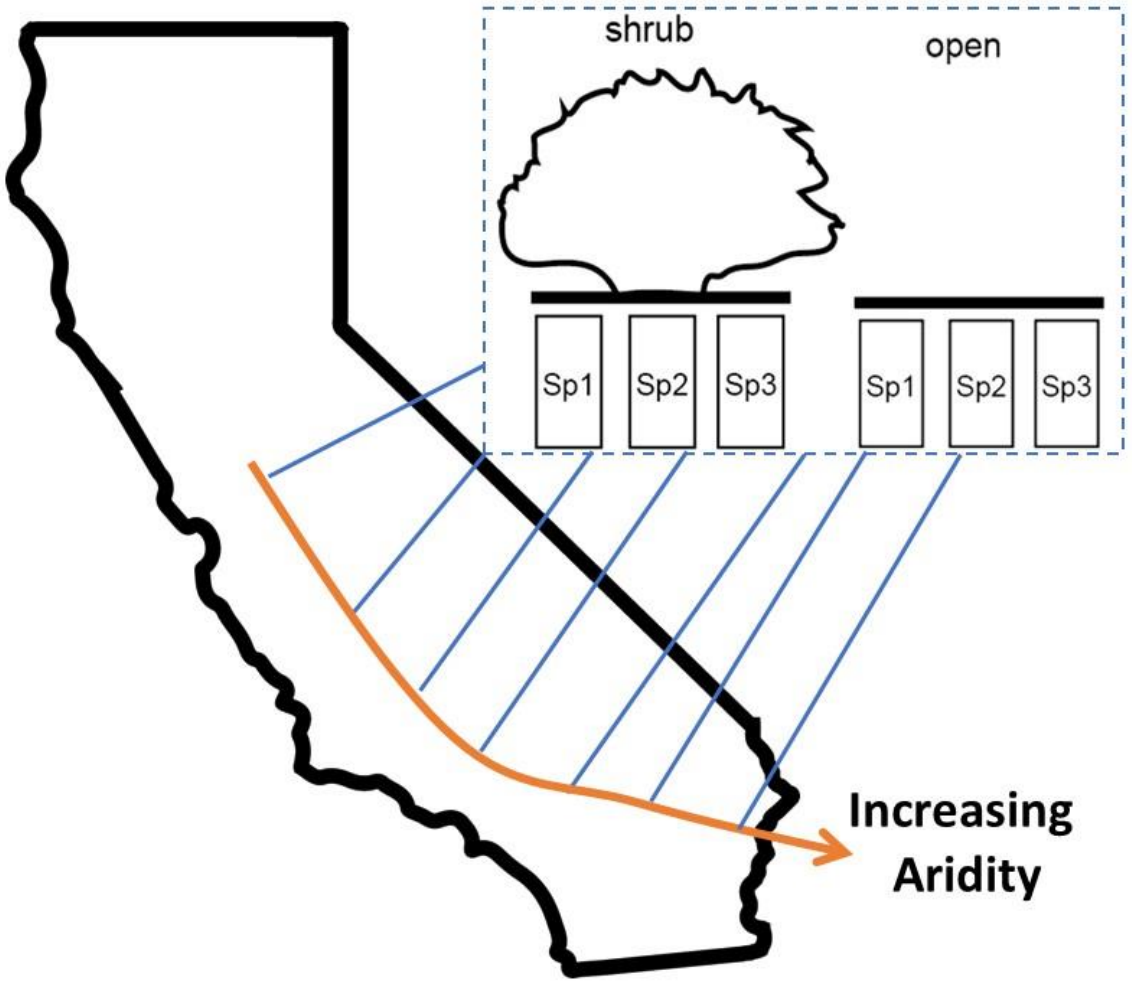
**ABSTRACT**

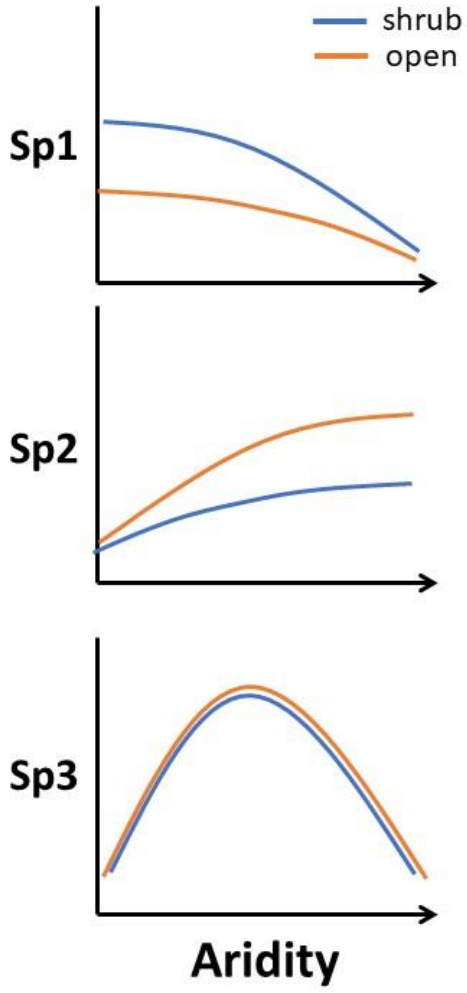
The use of linear mixed effects models (LMMs) is increasingly common in the analysis of biological data. Whilst LMMs offer a flexible approach to modelling a broad range of data types, ecological data are often complex and require complex model structures, and the fitting and interpretation of such models is not always straightforward. The ability to achieve robust biological inference requires that practitioners know how and when to apply these tools. Here, we provide a general overview of current methods for the application of LMMs to biological data, and

https://peerj.com/articles/4794/

# Case Study: Mixed models to estimate plant responses to aridity

# Aridity gradient to test plant interactions




**Conducted in 2016**

**Conducted in 2017**

*Filazzola et al. 2020*

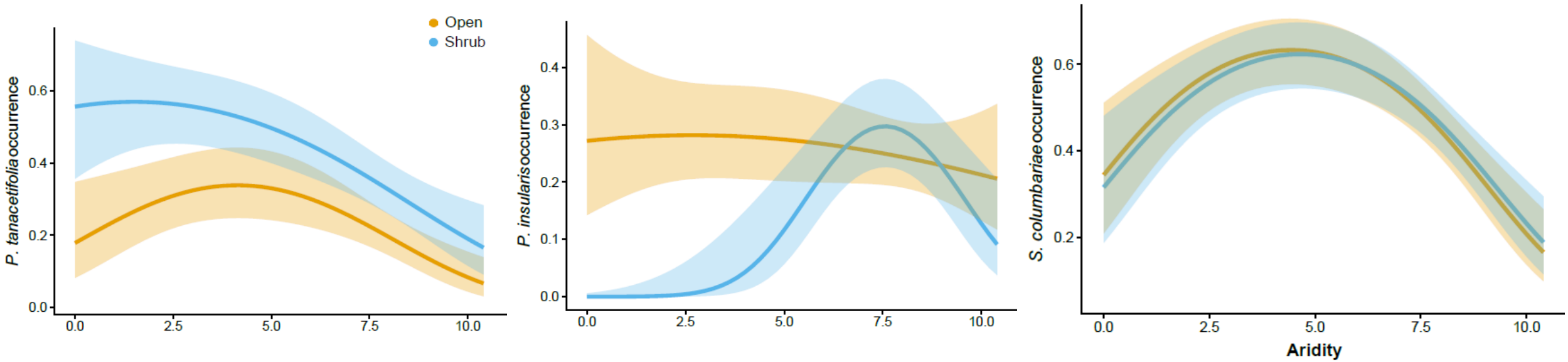
# Data of plant responses to aridity along gradient

| PlotID | Year | Site | Aridity | Microsite | Phacelia | Plantago | Salvia |
|---|---|---|---|---|---|---|---|
| 1 | 2016 | Barstow | 4.23 | open | 0 | 0 | 1 |
| 1 | 2016 | Barstow | 4.23 | shrub | 0 | 1 | 1 |
| 2 | 2016 | Barstow | 4.23 | open | 1 | 0 | 0 |
| 2 | 2016 | Barstow | 4.23 | shrub | 0 | 0 | 0 |
| 3 | 2016 | Barstow | 4.23 | open | 0 | 0 | 1 |
| 3 | 2016 | Barstow | 4.23 | shrub | 0 | 0 | 0 |
| 4 | 2016 | Barstow | 4.23 | open | 1 | 0 | 1 |
| 4 | 2016 | Barstow | 4.23 | shrub | 0 | 0 | 0 |
| 5 | 2016 | Barstow | 4.23 | open | 0 | 0 | 0 |
| 5 | 2016 | Barstow | 4.23 | shrub | 0 | 1 | 1 |
| 6 | 2016 | Barstow | 4.23 | open | 0 | 1 | 1 |

*Filazzola et al. 2020*

# Model formulation

glmer(plantOcc ~ aridity * microsite + (1|Year), data=data, family="binomial")




*Filazzola et al. 2020*

UNIVERSITY OF TORONTO

# Thank you!

https://www.filazzola.info/

https://afilazzola.github.io/IntroGLM/