Investigation of methods to predict groundwater redox status

Scott Wilson, Mike Friedel, Murray Close, Phil Abraham, Laura Banasiak

Aims for this project

This project is part of “Our Land & Water” National Science Challenge

- Our challenge was to develop a national coverage of groundwater redox status to assist management of land & water resources and to contribute to national scale modelling of effects

Our aim for this initial part of the project was to develop the best possible predictive model of redox status, to make robust predictions

- ➢ Needs to be applied to areas with sparse WQ data

Why is groundwater redox status important?

- A key groundwater contaminant in NZ is nitrate
- Increased land use intensity is increasing groundwater nitrate levels, leading to adverse impacts on lakes and lowland streams
- Only permanent removal process for N is denitrification
- Groundwater redox status is the key factor which determines whether denitrification will take place within a particular area of a groundwater system

Sample redox status assignment

- Classify sample redox status using NO_3, Mn, Fe, SO_4 and DO

  modified system of McMahon & Chapelle (2008)

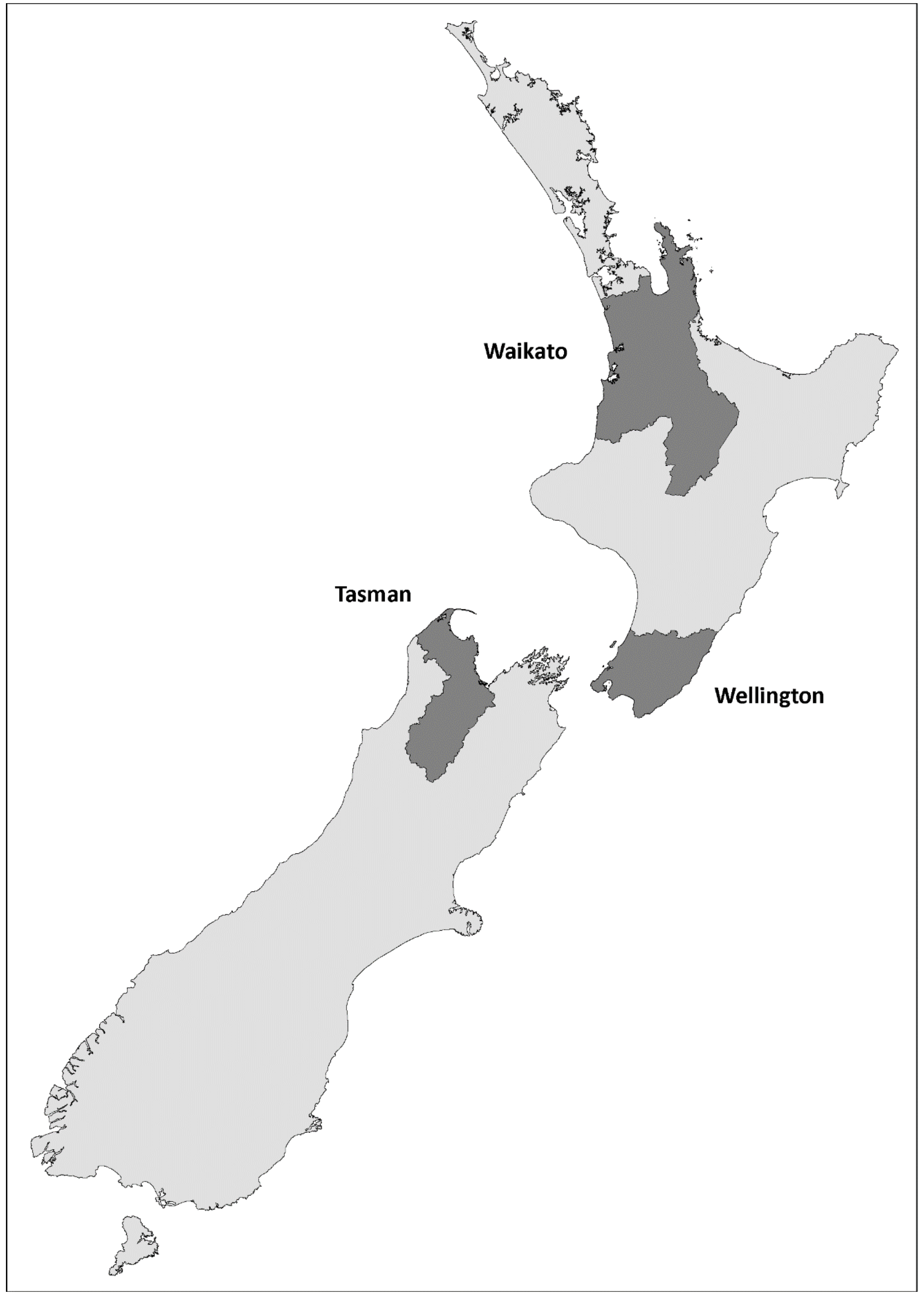
- 3 redox classes: Oxic Reduced Mixed

  - Oxic → High NO_3, SO_4, DO; Low Mn, Fe
  - Reduced → Low NO_3, SO_4, DO; High Mn, Fe
  - Mixed → High Mn; Low NO_3, Fe classed as mixed

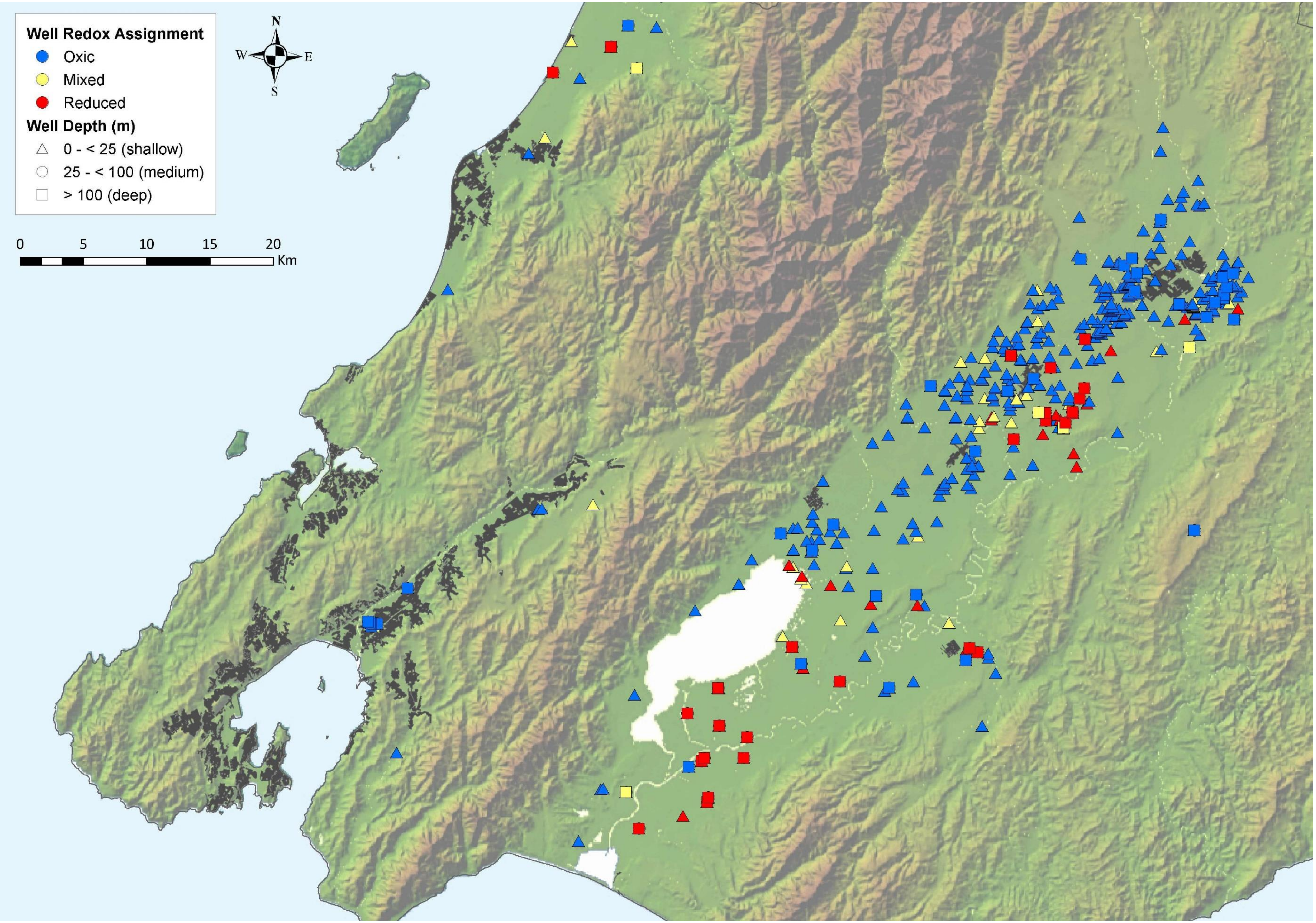
- Approach previously applied to Waikato, Canterbury, Southland

- Regions in current study are Waikato, Wellington, Tasman

Study Areas



	Tasman		Wellington		Waikato	
	Sampled	**%**	**Sampled**	**%**	**Sampled**	**%**
Oxic	598	86.8	376	81.2	375	63.8
Mixed	39	5.7	38	8.2	105	17.9
Reduced	52	7.5	49	10.6	108	18.4
Wells Sampled	**689**		**463**		**588**	



E/S/R
Science for Communities

LINCOLN AGRITECH LTD

Spatial Attributes

Spatial Attribute	Mapped Scale	Data Source	Reference
Groundwater depth	1000m raster	GNS: supplied	Westerhoff et al. (2018)
Land surface recharge	1000m raster		Westerhoff (2017)
Main rock Sub rock Geological age	1:50 000	GNS: QMap	Rattenbury and Heron (1997)
Soil order Soil drainage	1:50 000	Landcare: SMap & Fundamental Soil Layer	Hewitt (2010), Lilburne et al. (2012) Newsome et al. (2008)
Soil C_{max} Soil C_{min}	1:63 360	Landcare: NZ Fundamental Soil Layer	Newsome et al. (2008)
Rainfall PET AET	500m raster	NIWA: supplied	Tait & Woods (2007) Woods et al. (2006)
Mean annual low flow Mean flow February flow Fre3 flow	500m raster	MfE: data generated by NIWA	Snelder & Biggs (2002) Booker (2013 & 2015)
Landuse	1:50 000	Landcare: LUCAS NZ Land Use Map 2012	Newsome et al. (2013)
Nitrogen leaching	100m raster	MfE: data generated by AgResearch	Dymond et al. (2013)
Elevation Land slope	8m raster	Geographx 8m DEM	Geographx (2012)

- Discrete data
- Continuous data

Spatial Attributes

Spatial Attribute	Mapped Scale	Data Source	Reference
Groundwater depth	1000m raster	GNS: supplied	Westerhoff et al. (2018)
Land surface recharge	1000m raster		Westerhoff (2017)
Main rock Sub rock Geological age	1:50 000	GNS: QMap	Rattenbury and Heron (1997)
Soil order Soil drainage	1:50 000	Landcare: SMap & Fundamental Soil Layer	Hewitt (2010), Lilburne et al. (2012) Newsome et al. (2008)
Soil C_{max} Soil C_{min}	1:63 360	Landcare: NZ Fundamental Soil Layer	Newsome et al. (2008)
Rainfall PET AET	500m raster	NIWA: supplied	Tait & Woods (2007) Woods et al. (2006)
Mean annual low flow Mean flow February flow Fre3 flow	500m raster	MfE: data generated by NIWA	Snelder & Biggs (2002) Booker (2013 & 2015)
Landuse	1:50 000	Landcare: LUCAS NZ Land Use Map 2012	Newsome et al. (2013)
Nitrogen leaching	100m raster	MfE: data generated by AgResearch	Dymond et al. (2013)
Elevation Land slope	8m raster	Geographx 8m DEM	Geographx (2012)

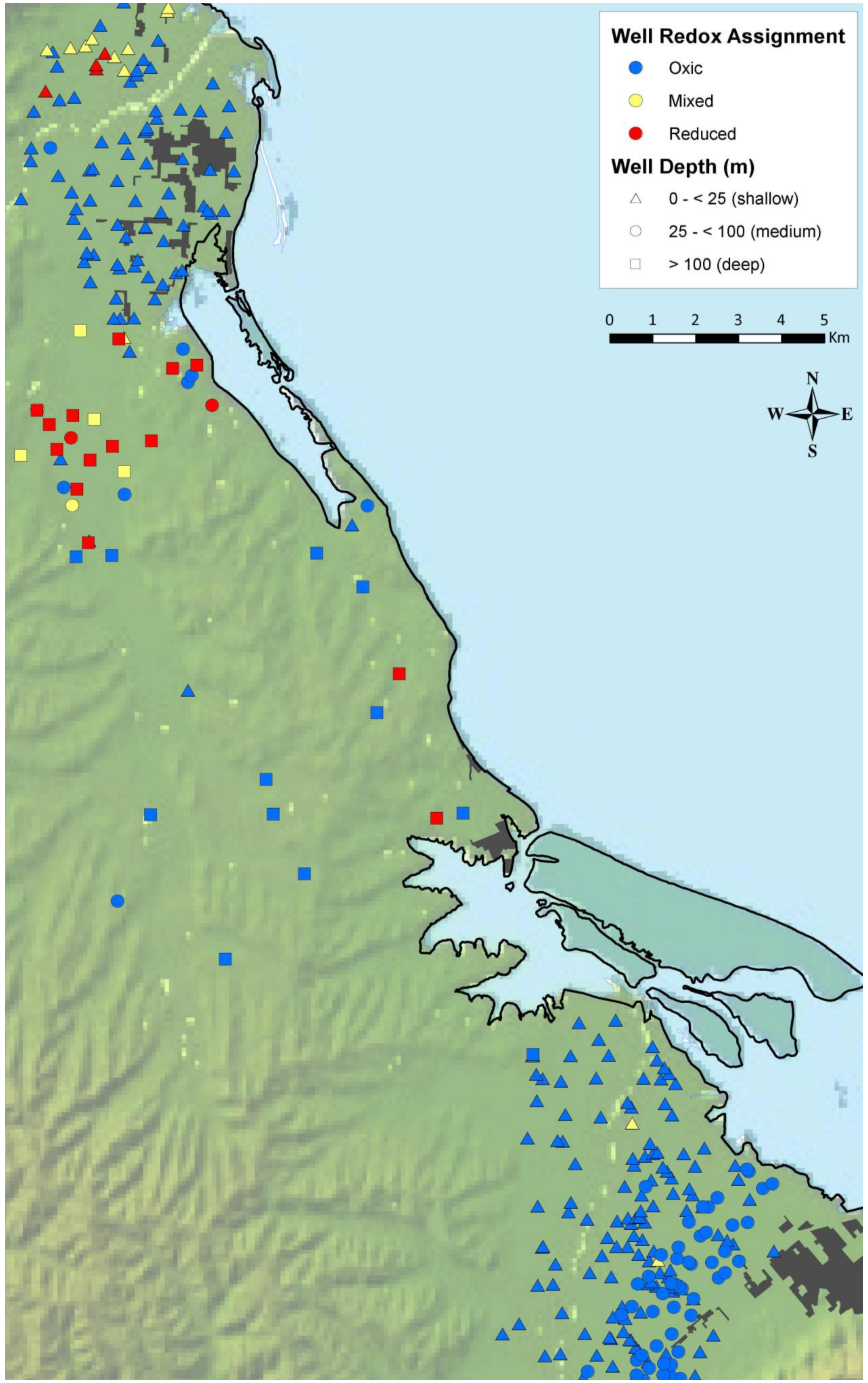
ESR Science for Communities

Initial 22 reduced to 14 using correlation matrix

LINCOLN AGRITECH LTD

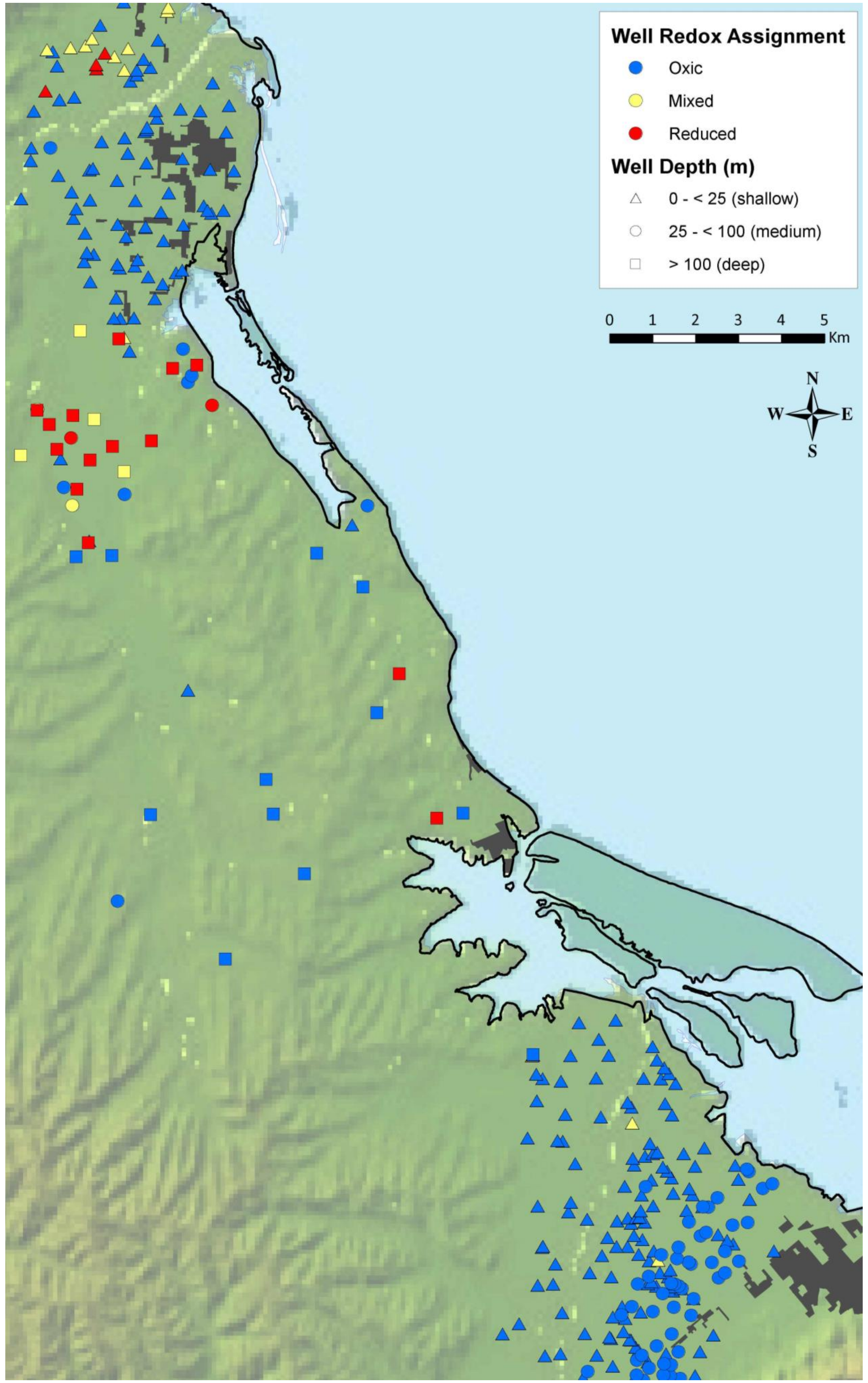
Prediction approach

- Redox assignment (response variable) ↔ spatial attributes (predictors)
- Previously used Linear Discriminant Analysis (LDA)
    - Close, M.; Abraham, P.; Humphries, B.; Lilburne, L.; Cuthill, T.; Wilson S. 2016. Predicting Groundwater Redox Status on a Regional Scale using Linear Discriminant Analysis. *Journal of Contaminant Hydrology 191*: 19–32.
    - Wilson, S., Close, M., Abraham, P., 2018. Applying Linear Discriminant Analysis to predict groundwater redox conditions conducive to denitrification. *Journal of Hydrology 556*: 611-624.
- For this study we compared predictions using LDA to those using nonlinear methods - Random Forest (RF) & Boosted Regression trees
- No significant improvement in the solutions!
    - ➢ But all models were strongly influenced by data bias



Sources of bias

- Spatial bias (clustering)
- Depth bias (predominantly shallow)
- Sample selection bias (65-85% oxic)
- Attribute bias
  - ➢ Samples unevenly distributed among spatial attribute categories
  - ➢ Sampling <1 % of the unique attribute combinations



Sources of bias

- Spatial bias (clustering)
- Depth bias (predominantly shallow)
- **Sample selection bias (65-85% oxic)**
- **Attribute bias**
  - **Samples unevenly distributed among attribute categories**
  - **Sampling <1 % of the unique attribute combinations**

Sample Selection Bias

Attribute bias removed from the RF model using cForest

- but still had massive issues with the bias from sample selection bias

Skewness in the distribution of WQ data – predominantly sampling oxic groundwater

If you have 80% oxic water, you get ~80% predictive accuracy from the model due to dominance of one class of response variable

- but not a good model (null case test)
- Cohen's Kappa metric – gives model predictive power taking random agreement into account (possibility of agreement due to chance)

Sample Selection Bias

Model	Tasman (84% Oxic)		Wellington (80% Oxic)		Waikato (65% Oxic)	
	Accuracy	Kappa	Accuracy	Kappa	Accuracy	Kappa
cForest (null)	0.84	0	0.80	0	0.65	0
cForest (Attrib. bias adjusted)	0.84	0.14	0.81	0.13	0.66	0.10
LDA	0.87	0.28	0.84	0.34	0.67	0.22

- Model predictive agreement is slight to fair (perfect agreement =1)
- Randomly deleting response variable data to improve proportions increases Kappa, but decreases accuracy

Hybrid Machine Learning workflow

E/S/R Science for Communities

1. Development

Unsupervised learning - mapping & self-organization of redox & predictor variables on hypersurface (75% data)

Metaheuristics - selection of predictor variables by supervised machine learning (KNN) into genetic algorithm

Performance metrics - Kappa statistic (class), cross-validation (continuous)

Iterative Process

2. Generalisation

- **Prediction** - simultaneous redox probability with holdout predictor variables (25% data)
- **Performance metrics** – Kappa statistic (class), cross-validation (continuous)

3. Prediction

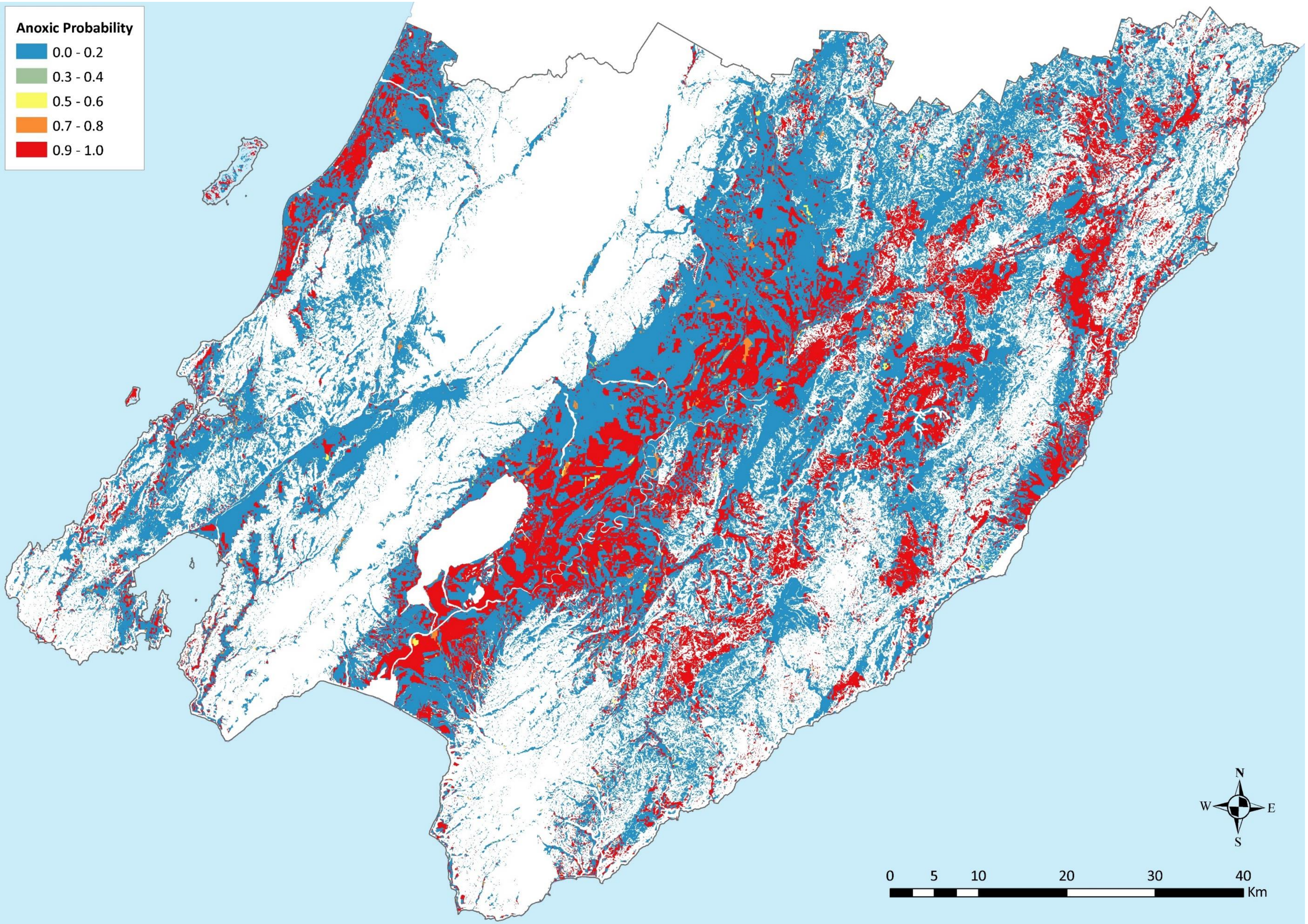
Simultaneous redox probability with independent predictor variables at 130,000+ locations

LINCOLN AGRITECH LTD

ML Model performance

Hybrid model performs superbly for both accuracy and kappa metrics

Model	Tasman (84% Oxic)		Wellington (80% Oxic)		Waikato (65% Oxic)	
	Accuracy	Kappa	Accuracy	Kappa	Accuracy	Kappa
cForest (null)	0.84	0	0.80	0	0.65	0
cForest (Attrib. bias adjusted)	0.84	0.14	0.81	0.13	0.66	0.10
LDA	0.87	0.28	0.84	0.34	0.67	0.22
Hybrid	1.0	1.0	0.92	0.98	0.76	0.87



ESR Science for Communities

LINCOLN AGRITECH LTD

Conclusions

- Beware the effect of sample selection bias on statistical model development!

- Prediction for attribute combinations outside model range can be very low
  - ➢ Significant issue as we extend our predictions to national coverage

- New Hybrid ML approach successfully overcomes these sources of bias

- Next steps:
  - ➢ Regionalise the data (group areas of similarity & move away from council boundaries)
  - ➢ Apply the approach to these regions to achieve a national coverage of regional scale maps