Investigation of methods to predict groundwater redox status with variable amounts of available well data

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

Outline

- Background and Aims
- Previous method for prediction of groundwater redox status
  - Issues with bias
  - Issues with sparse data
- Case study areas – Waikato, Wellington, Tasman
- Revised method for prediction of groundwater redox status
- Results & Conclusions

Why is groundwater redox status important?

- Key groundwater contaminant in NZ is nitrate
- Increased land use intensity is increasing groundwater nitrate levels, leading to impacts on lakes and lowland streams
- Only permanent removal process for N is denitrification
- Groundwater redox status is the key factor whether denitrification will take place in a particular part of a groundwater system
- Would like to be able to predict redox status for all groundwater in a catchment, not just at wells

Aims for this project

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

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

- Significant variability across NZ in terms of GW resources and availability of GW quality data

Aims for this part of the project were to develop robust predictions of redox status that:

- Accounted for bias
- Could be used in areas with variable amounts of WQ data

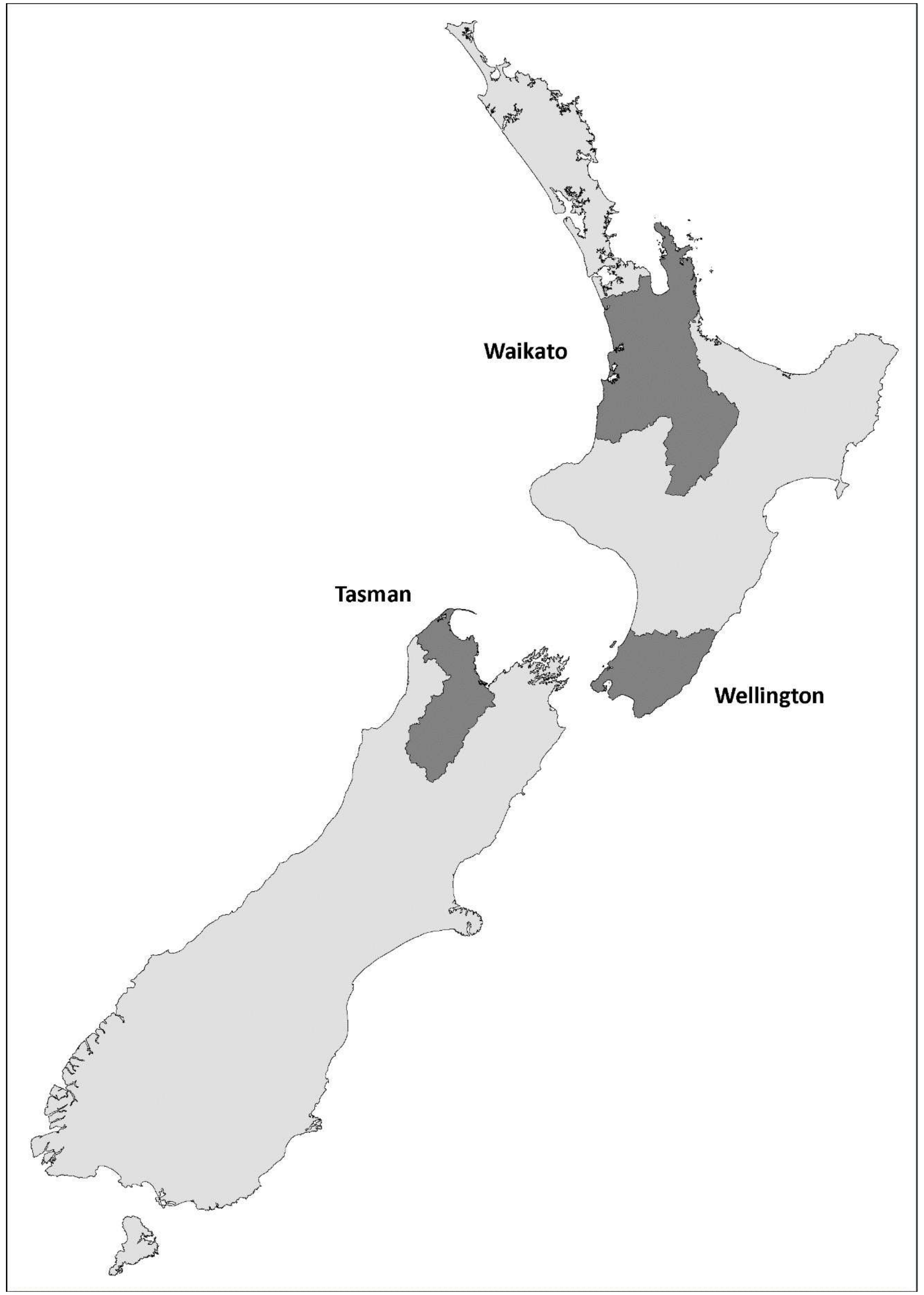
Redox status assignment

E/S/R Science for Communities

- Classify each well's redox status using NO_3, Mn, Fe, SO_4 and DO using modified protocol from McMahon & Chapelle (2008); each parameter was classed as high or low depending on a threshold value.
- Used 3 redox classes: Oxic, Mixed and Reduced
  - High NO_3, SO_4, DO; Low Mn, Fe classed as oxic
  - Low NO_3, SO_4, DO; High Mn, Fe classed as reduced
  - High Mn; Low NO_3, Fe classed as mixed
- Previously applied to Waikato, Canterbury and Southland
- Regions in current study are Waikato, Wellington and Tasman

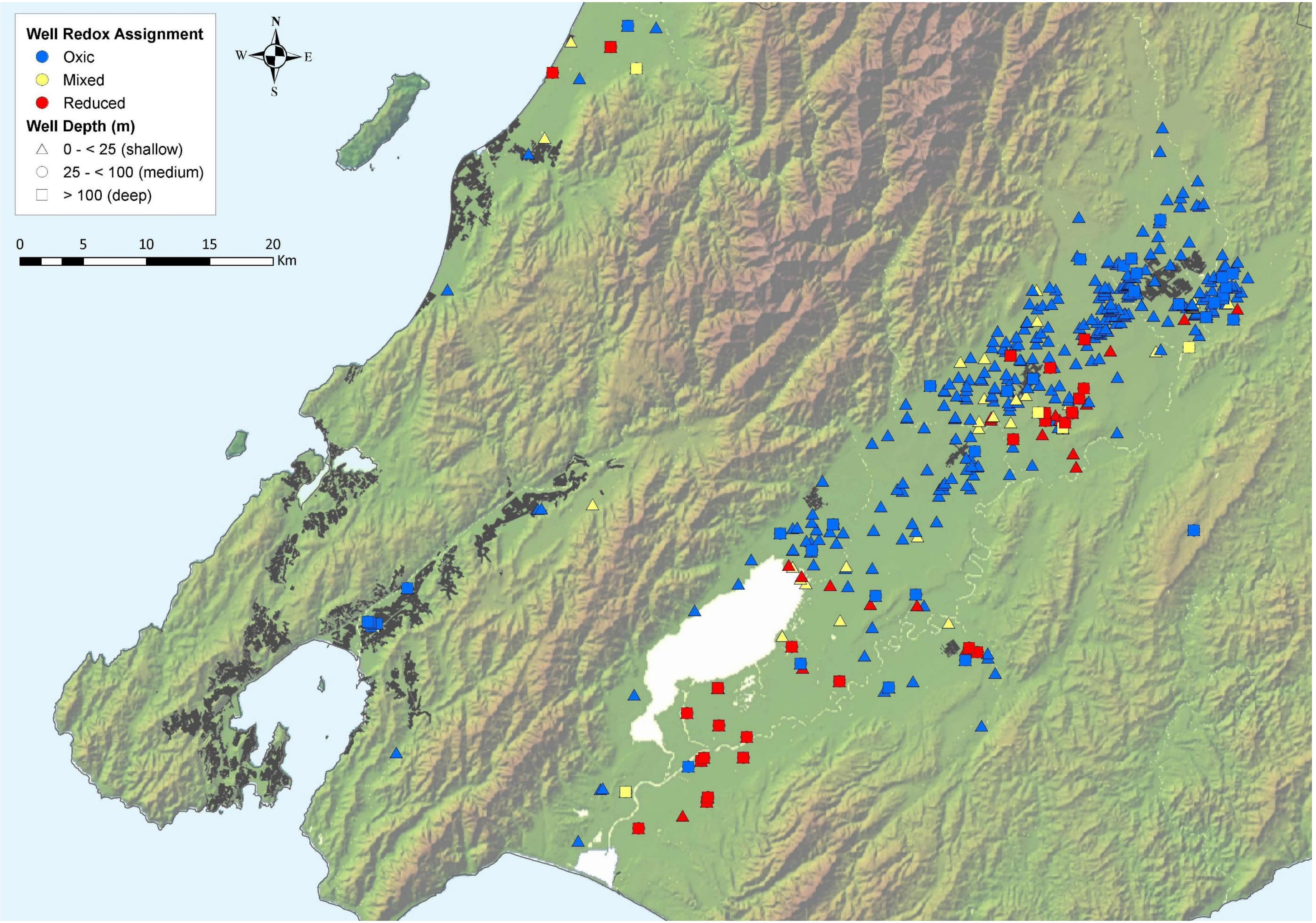
LINCOLN AGRITECH LTD

Study Areas



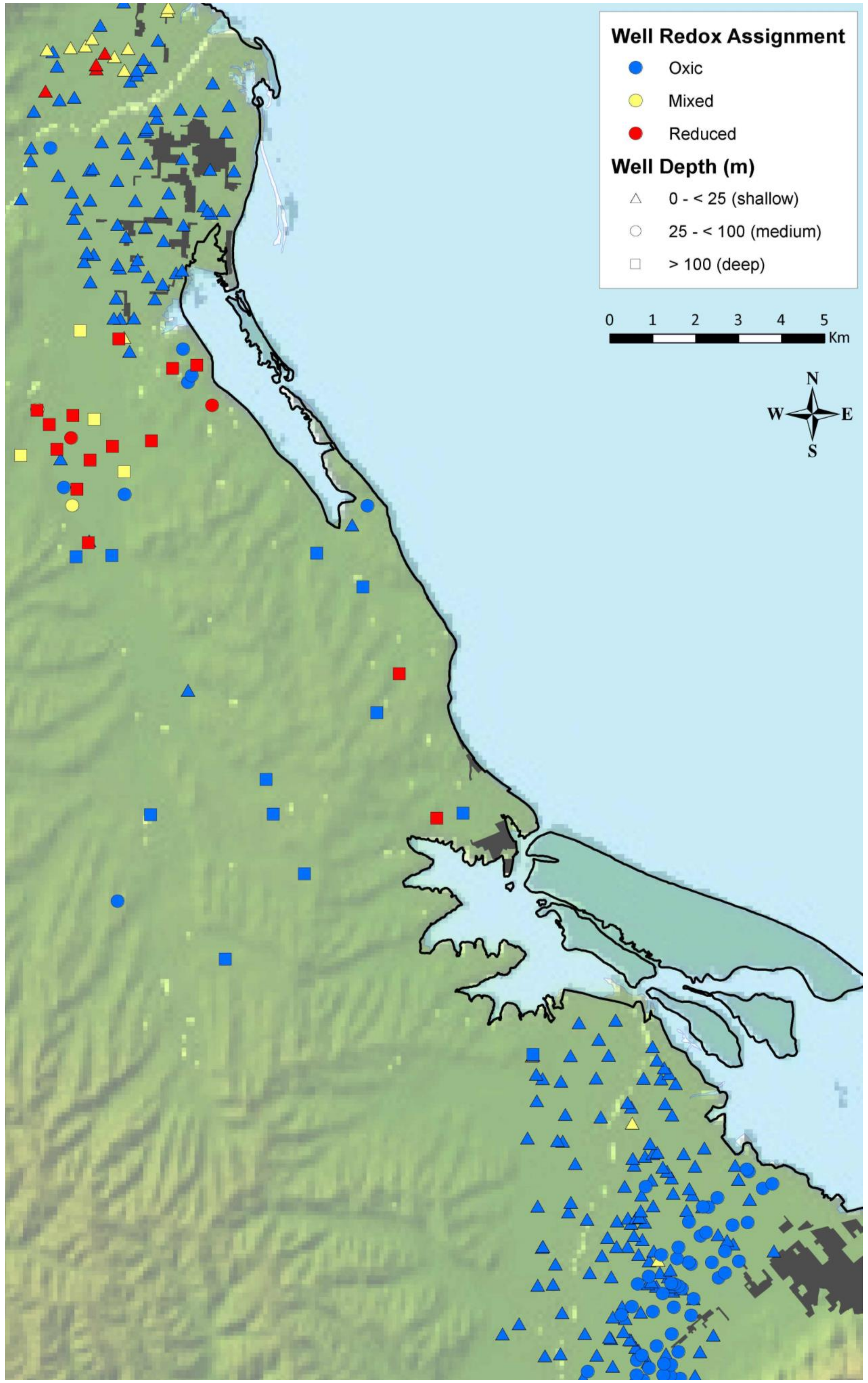
	Tasman		Wellington		Waikato	
	Sampled	%	Sampled	%	Sampled	%
Oxic	598	86.8	376	81.2	375	63.8
Mixed	39	5.7	38	8.2	105	17.9
Reduced	52	7.5	49	10.6	108	18.4
Total	**689**		**463**		**588**	

Redox status assignment



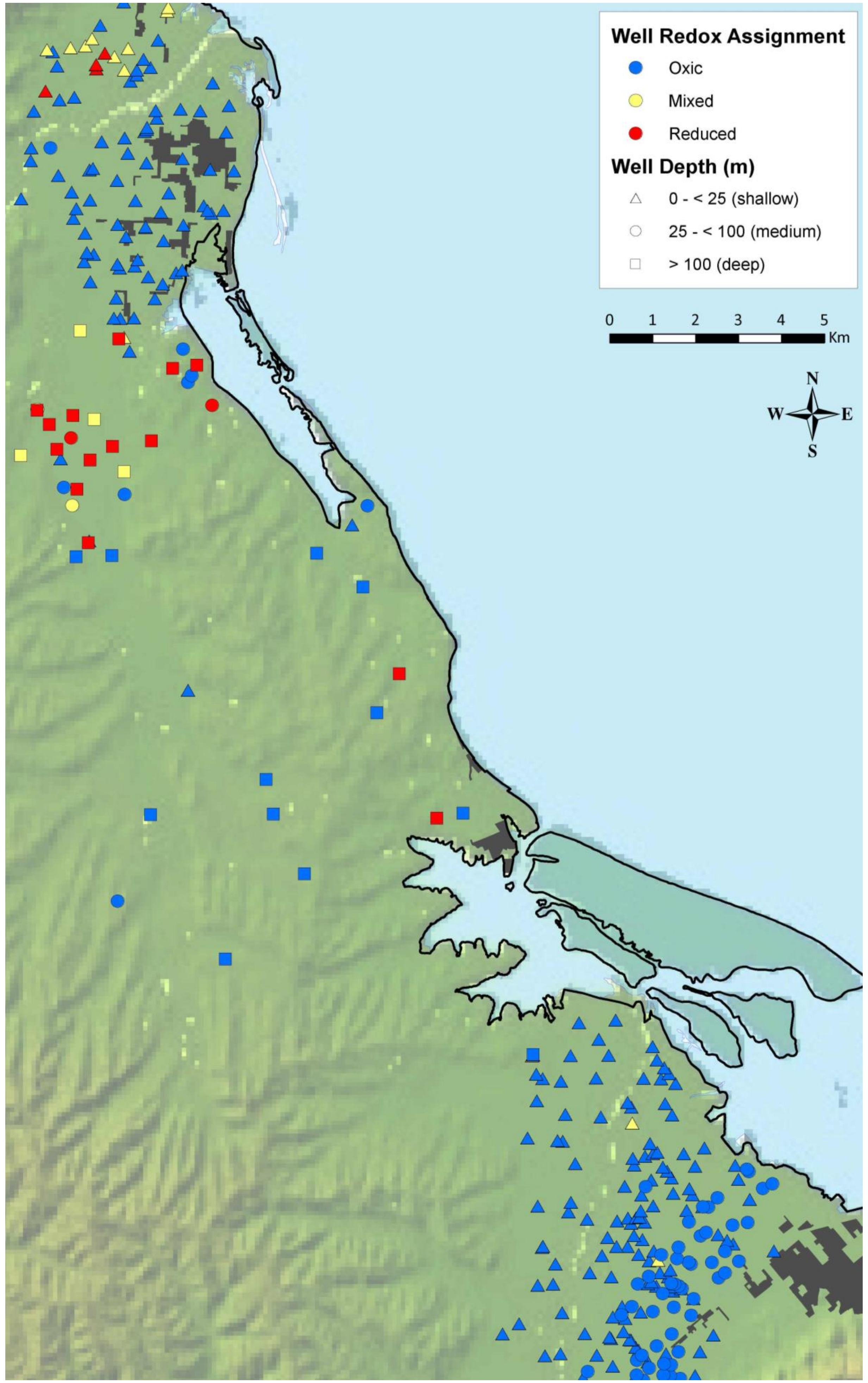
Redox prediction using Linear Discriminant Analysis (LDA)

- Used LDA to predict redox status based on a range of geological, land use; topography, soil, and hydrological parameters
- This approach has been presented previously and published
  - 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.
- We compared the predictions using LDA to those using Random Forest (RF)
- Both approaches gave reasonable solutions but as we tried to develop models for regions with less available groundwater quality, we encountered 2 issues: Bias and Sparse data



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 0.25-0.5 % 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 0.25-0.5 % of the unique attribute combinations**

Predictive Attributes

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

Sample Selection Bias

Bias from Skewness in the distribution of WQ data – most well samples indicated oxic groundwater

If you had 80% oxic water, then if you predicted everything as oxic => 80% accuracy; but not a good model – shown as null case in next slide

- Kappa metric – gives model predictive power taking random agreement and sample selection bias into account
- We removed the attribute bias from the RF model but still had massive issues with the bias from sample skewness

Sample Selection Bias

Bias from Skewness in the distribution of WQ data – most well samples indicated oxic groundwater

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

Overview of hybrid Machine Learning approach

1. Development

Unsupervised learning - mapping and self-organization of redox, depth, predictor variables on hypersurface (75% data at 150 locations)

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

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

Iterative Process

2. Generalisation

- **Prediction** - simultaneous redox probability and depth with holdout predictor variables (25% data at 150 locations)
- **Performance metrics** – Kappa statistic (class), cross-validation (continuous)

3. Prediction

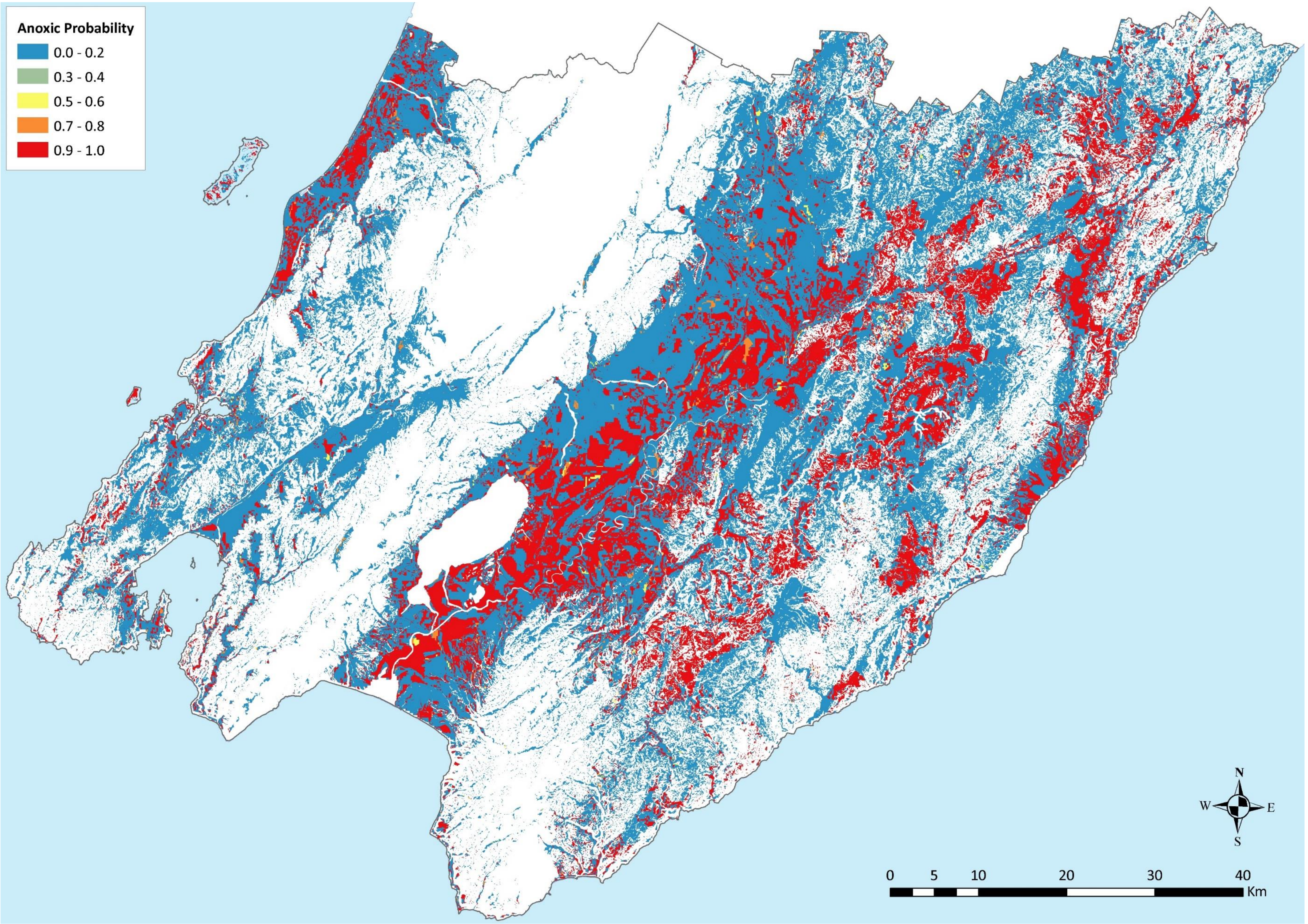
Simultaneous redox probability and depth with independent predictor variables at 130,000+ locations

Model performance

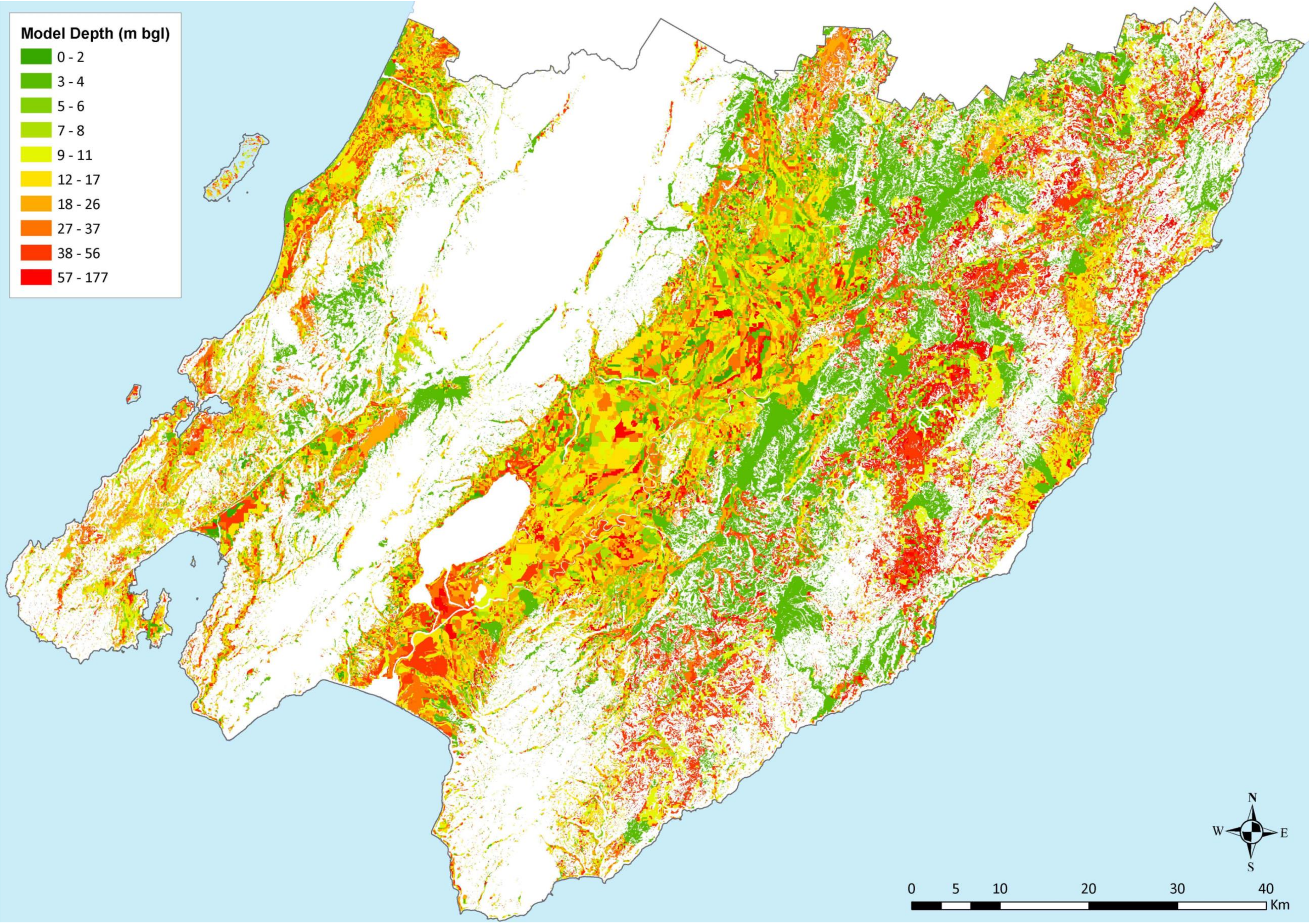
Hybrid model much better for both accuracy and kappa metrics

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

Model Results: anoxic probability



Model Results: prediction depths



Science for Communities

Conclusions

- Bias can be important issue
- Accuracy can be high but prediction outside model input range can be very low – significant issue as we tried to extend our predictions to national coverage
- New Hybrid Machine Learning approach overcomes these sources of bias
- Next steps are to develop new regional grouping (not use RC boundaries) and to apply the approach to more regions to achieve a national coverage of regional scale maps

Any Questions???

Murray Close

E: murray.close@esr.cri.nz