Australian Government

Geoscience Australia

Novel hybrid spatial predictive methods of machine learning and geostatistics with applications to terrestrial and marine environments in Australia

Jin Li* and Augusto Sanabria

National Earth & Marine Observations

Environmental Geoscience Division

Geoscience Australia

* jin.li@ga.gov.au

APPLYING GEOSCIENCE TO AUSTRALIA'S MOST IMPORTANT CHALLENGES

Datasets, maps & comments

Bob Cechet

Ian French

Riko Hashimoto

Zhi Huang

Chris Lawson

Xiaojing Li

Tony Nicholas

Scott Nichol

Daniel McIlroy

Anna Potter

Xuerong Qin

Tanya Whiteway

Functions in sp, gstat and raster packages in R

Roger Bivand (Norwegian School of Economics and Business Administration),

Paul Hiemstra (University of Utrecht),

Robert Hijmans (University of California),

Edzer Pebesma (University of Münster),

Michael Summer (University of Tasmania).

Australian Government
Geoscience Australia

Comparison of spatial interpolation methods using a simulation experiment based on Australian seabed sediment data

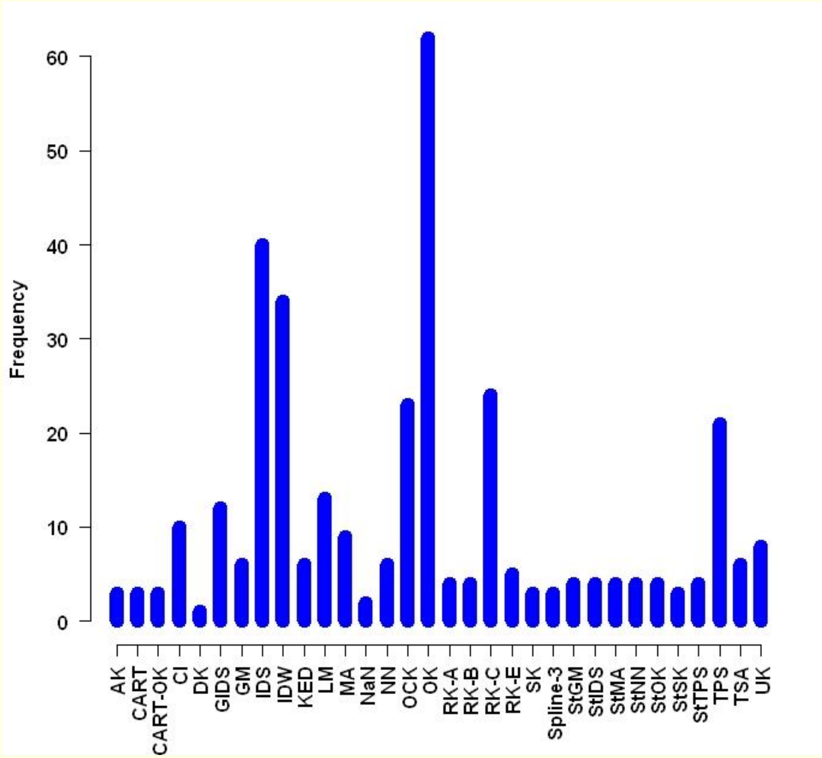
Jin Li*, Andrew Heap, Anna Potter & James Daniell

Marine & Coastal Environment
* jin.li@ga.gov.au

useR!

GEOSCIENCE AUSTRALIA

Classification of the existing methods (Li and Heap 2008):

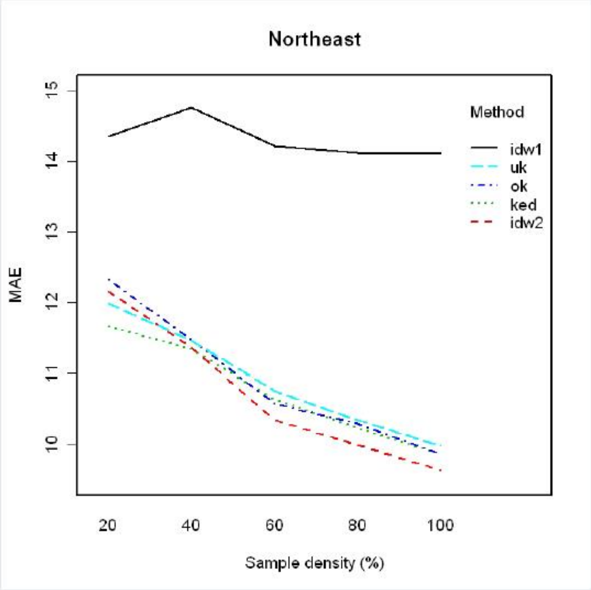


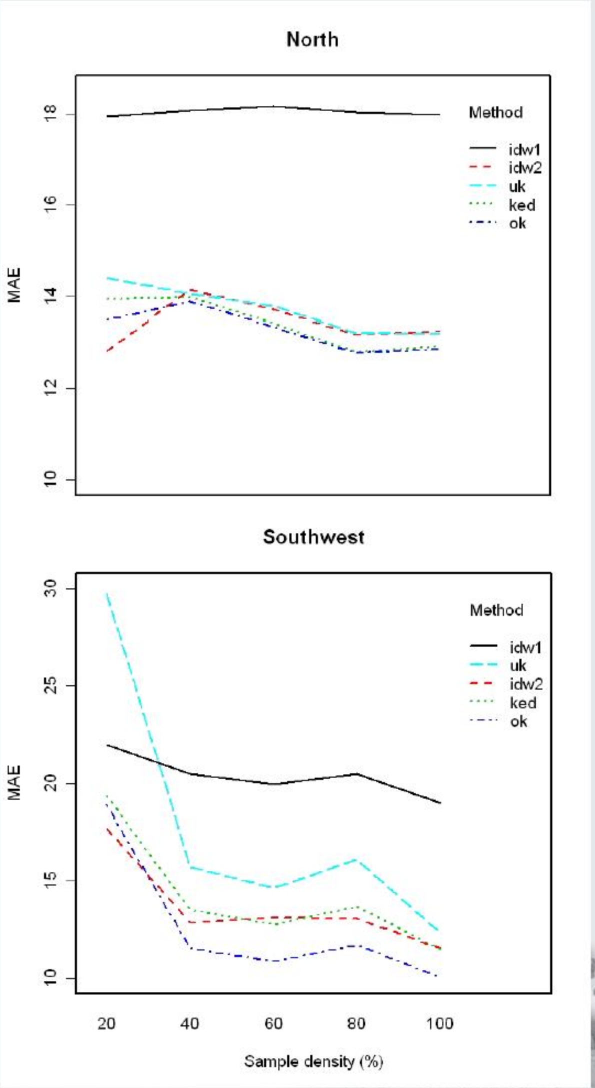
AK	Akima's interpolator
CART	regression tree
CI	classification
DK	disjunctive kriging
GIDS	gradient plus IDS
GM	global mean
IDS	inverse distance squared
IDW	inverse distance weighting
KED	kriging with an external drift
LM	linear regression model
MA	moving average
NaN	natural neighbours
NN	nearest neighbours
OCK	ordinary CK
OK	ordinary kriging
RK	regression kriging
SK	simple kriging
Spline-3	cubic spline
TPS	thin plate splines
TSA	trend surface analysis
UK	universal kriging

The frequency of 32 spatial interpolation methods compared in 80 cases (Li & Heap, 2008 and 2011).

- Non-geostatistical methods (e.g., inverse distance squared: IDS)
- Geostatistical methods (e.g., ordinary kriging: OK)
- Combined methods (e.g. regression kriging: RK)

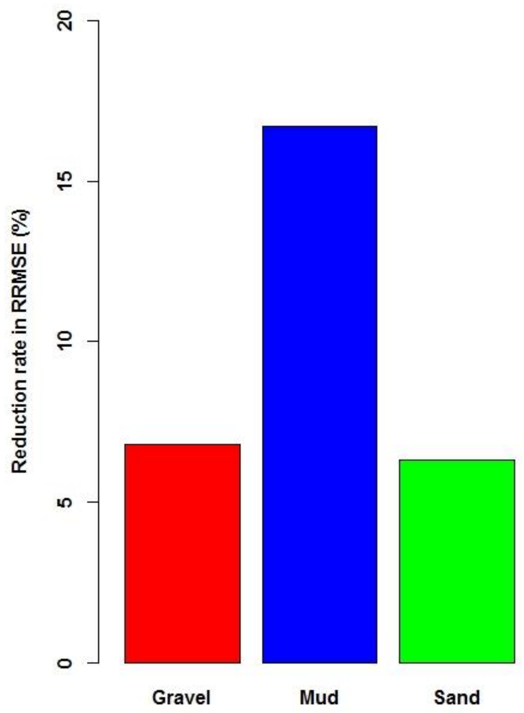
Interaction among sample density, method and region





1	Inverse distance weighting (IDW)
2	Generalised least squares trend estimation (GLS)
3	Kriging with an external drift (KED)
4	Ordinary cokriging (OCK)
5	Ordinary kriging (OK)
6	Universal kriging (UK)
7	Boosted regression tree (BRT)
8	General Regression Neural Network (GRNN)
9	RandomForest (RF)
10	Regression tree (RT)
11	Support vector machine (SVM)
12	Thin plate splines (TPS)
13	Linear models and OK (RKlm)
14	Generalised linear models and OK (RKglm)
15	Generalised least squares and OK (RKgls)
16	BRT and OK (BRTOK)
17	BRT and IDS (BRTIDS)
18	GRNN and OK (GRNNOK)
19	GRNN and IDS (GRNNIDS)
20	RF and IDS (RKIDS)
21	RF and OK (RKOK)
22	RT and OK (RTOK)
23	RT and IDS (RTIDS)
24	SVM and OK (SVMOK)
25	SVM and OK (SVMIDS)

1	Inverse distance weighting (IDW)
2	Generalised least squares trend estimation (GLS)
3	Kriging with an external drift (KED)
4	Ordinary cokriging (OCK)
5	Ordinary kriging (OK)
6	Universal kriging (UK)
7	Boosted regression tree (BRT)
8	General Regression Neural Network (GRNN)
9	RandomForest (RF)
10	Regression tree (RT)
11	Support vector machine (SVM)
12	Thin plate splines (TPS)
13	Linear models and OK (RKlm)
14	Generalised linear models and OK (RKglm)
15	Generalised least squares and OK (RKgls)
16	BRT and OK (BRTOK)
17	BRT and IDS (BRTIDS)
18	GRNN and OK (GRNNOK)
19	GRNN and IDS (GRNNIDS)
20	RF and IDS (RKIDS)
21	RF and OK (RKOK)
22	RT and OK (RTOK)
23	RT and IDS (RTIDS)
24	SVM and OK (SVMOK)
25	SVM and OK (SVMIDS)



Reduction rate in predictive error (RRPE) by the hybrid methods of Machine Learning Methods and the Existing Spatial Predictive Methods (RF/RFOK/RFIDS) in comparison with IDS based on previous studies (Li et al. 2010, 2011a, b, c, and 2012).

RRMSE: relative root mean squared error.

RRPE = (PE_control - PE_tested)/PE_control*100
PE: predictive error.

Development of the Hybrid Methods of Machine Learning and the Existing Spatial Predictive Methods

No	Method
1	**the combination of random forest (RF) and OK (RFOK)**
2	**the combination of RF and IDS (RFIDS)**
3	the combination of support vector machine (SVM) and OK (SVMOK)
4	the combination of SVM and IDS (SVMIDS)
5	the combination of boosted regression tree (BRT, a version of gbm) and OK (BRTOK)
6	the combination of BRT and IDS (BRTIDS)
7	the combination of general regression neural network (GRNN) and OK (GRNNOK)
8	the combination of GRNN and IDS (GRNNIDS)

They were reviewed by Li & Heap (2014) and the first two methods were developed in 2008 at GA and published later (Li et al. 2010, Li 2011, Li et al. 2011a, b & c, Li et al. 2012, Li 2013a, b).

The superior performance of these hybrid methods was partially attributed to the features of RF, one component of the hybrid methods (Li et al. 2011b & 2011c).

One of the features is that RF selects the most important variable to split the samples at each node split for each individual trees, thus it is argued to implicitly perform variable selection (Okun and Priisalu, 2007). So the hybrids presumably also share this feature.

In this study we aim to address the following questions:

1) are they data-specific for marine environmental data?
2) is ‘model selection’ required for RF and the hybrid method? and
3) are these new hybrid methods equally applicable to terrestrial environmental data?

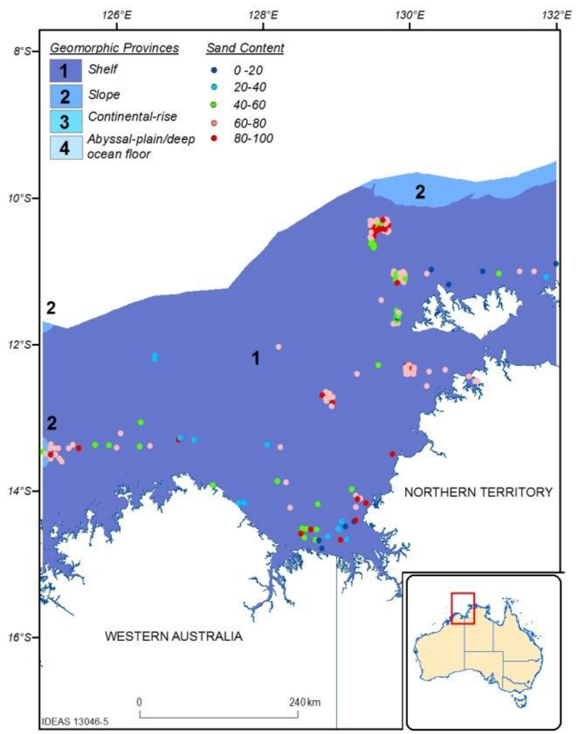
Application to Marine Environment

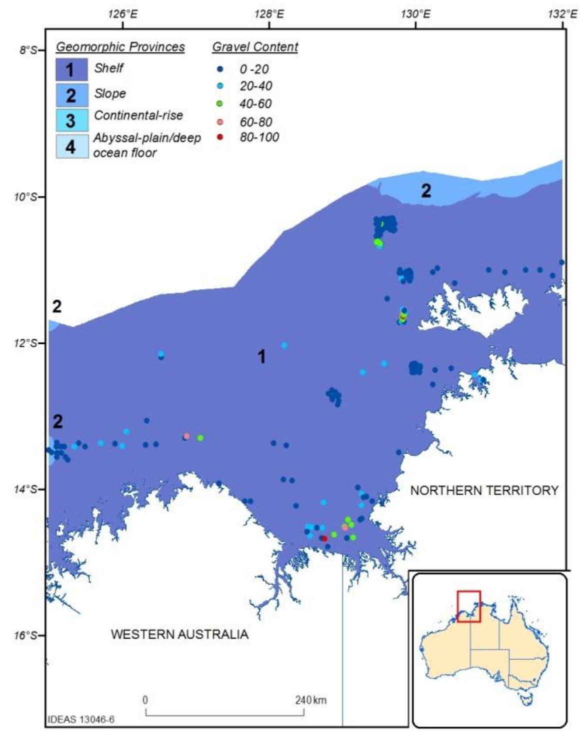
Region

Modelling methods

Accuracy assessment

<u>Sand</u> and <u>gravel</u> samples in the Timor Sea, Australia (*n*=238)





Application to Marine Environment

Region

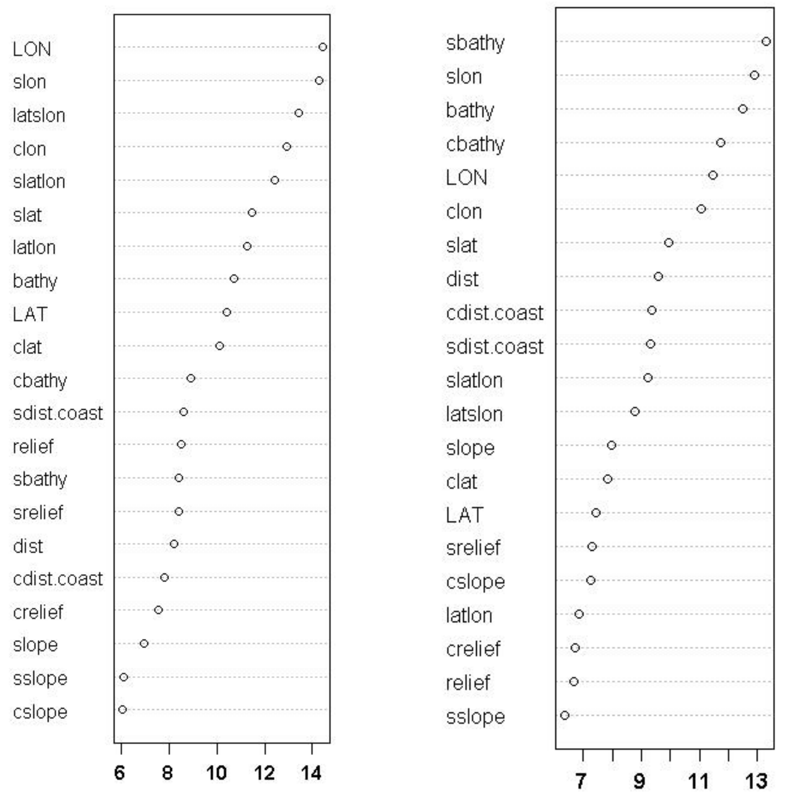
Modelling methods

Accuracy assessment

No	Method
1	IDW
2	OK
3	RFOK

Method	Predictive variables including **derived variables**
RFOK	**bathy, dist.coast, slope, relief, lat, long**, bathy^2, bathy^3, dist.coast^2, dist.coast^3, slope^2, slope^3, relief^2, relief^3, lat^2, long^2, lat*long, lat*long^2, long*lat^2, lat^3, long^3

Model selection: variable importance



Mean decrease in accuracy for sand & gravel content

Application to Marine Environment

Region

Modelling methods

Accuracy assessment

Performance of methods:

100 iterations of 10-fold cross-validation

Measures of predictive error (Li & Heap 2008 & 2011):

Relative mean absolute error (RMAE)

Relative root mean square error (RRMSE)

Reduction rate in predictive error (RRPE):

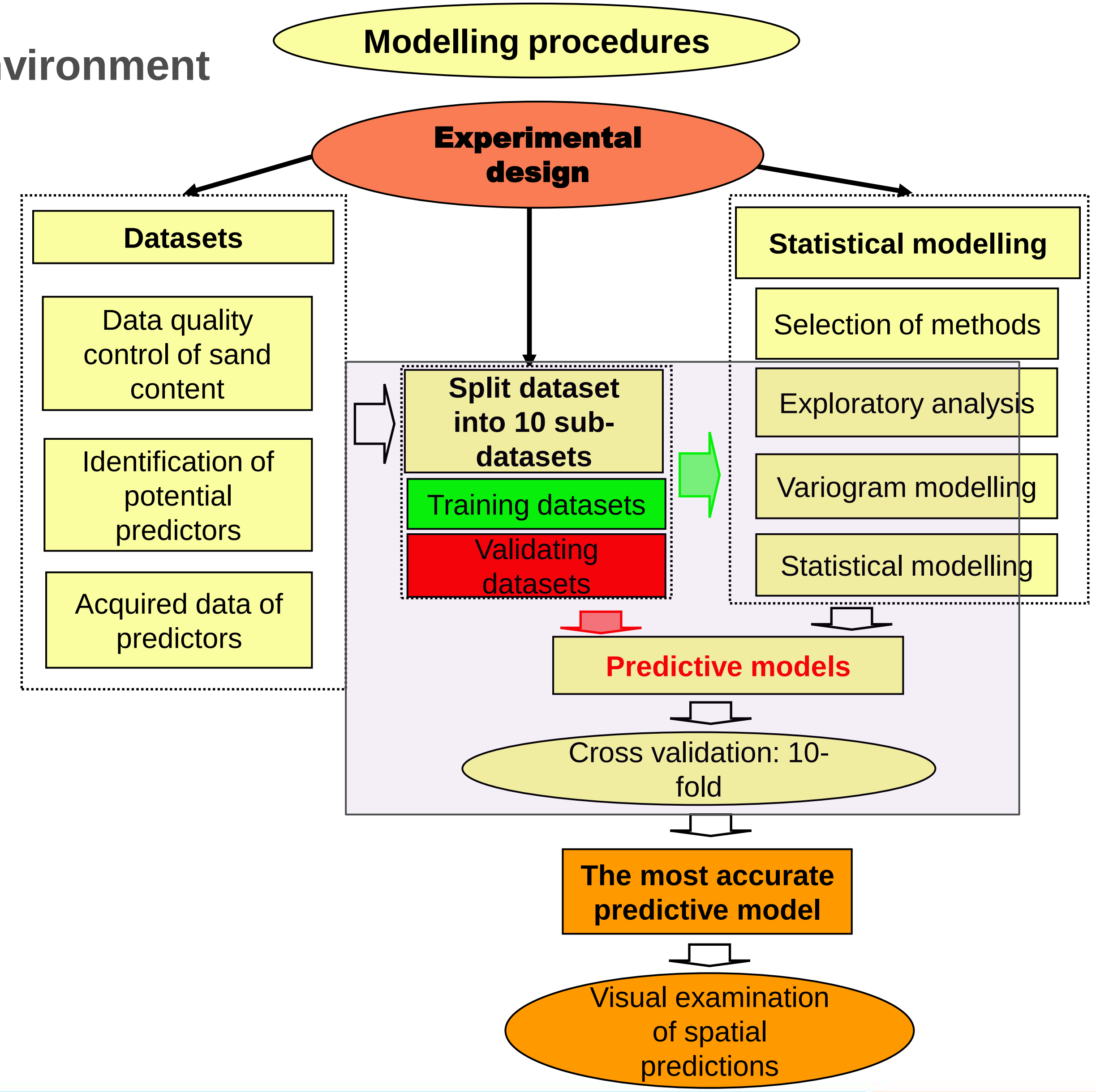
RRPE = (PE_control - PE_tested)/PE_control*100

PE: predictive error.

Software:

R 2.15.1

Application to Marine Environment



Application to Marine Environment

Effects of input variables

Sand content: 23 models

	Modelling.process	Predictors	No.predictors
1	Model 1: All 21 predictors	All 21 variables	21
2	Model 2: - cslope and sslope from model 1	lon, lat, bathy, dist, relief, slope, sbathy, cbathy, sdist.coast, cdist.coast, srelief, crelief, slat, clat, slon, clon, latlon, latslon, slatlon	19
3	Model 3: - slope from model 2	lon, lat, bathy, dist, relief, sbathy, cbathy, sdist.coast, cdist.coast, srelief, crelief, slat, clat, slon, clon, latlon, latslon, slatlon	18
4	Model 4: - srelief and cbathy from model 3	lon, lat, bathy, dist, relief, sbathy, sdist.coast, cdist.coast, crelief, slat, clat, slon, clon, latlon, latslon, slatlon	16
5	Model 5: - cdist.coast from model 4	lon, lat, bathy, dist, relief, sbathy, sdist.coast, crelief, slat, clat, slon, clon, latlon, latslon, slatlon	15
6	Modle 6: - sbathy from model 5	lon, lat, bathy, dist, relief, sdist.coast, crelief, slat, clat, slon, clon, latlon, latslon, slatlon	14
7	Model 7: - crelief from model 6	lon, lat, bathy, dist, relief, sdist.coast, slat, clat, slon, clon, latlon, latslon, slatlon	13
8	Model 8: - latlon from model 7	lon, lat, bathy, dist, relief, sdist.coast, slat, clat, slon, clon, latslon, slatlon	12
9	Model 9: - sdist.coast from model 8	lon, lat, bathy, dist, relief, slat, clat, slon, clon, latslon, slatlon	11
10	Modle 10: - relief from model 9	lon, lat, bathy, dist, slat, clat, slon, clon, latslon, slatlon	10
11	Model 11: - clat from model 10	lon, lat, bathy, dist, slat, slon, clon, latslon, slatlon	9
12	Model 12: - bathy from model 11	lon, lat, dist, slat, slon, clon, latslon, slatlon	8
13	Model 13: - dist from model 12	lon, lat, slat, slon, clon, latslon, slatlon	7
14	Model 14: - slatlon from model 13	lon, lat, slat, slon, clon, latslon	6
15	Model 15: - clon from model 14	lon, lat, slat, slon, latslon	5
16	Modle 16: - slat from model 15	lon, lat, slon, latslon	4
17	Model 17: - slon from model 16	lon, lat, latslon	3
18	Model 18: - latslon from model 17	lon, lat	2
19	Model 19: - lon from model 18	lat	1
20	Model 20: lon, lat, bathy, dist, relief, slope	lon, lat, bathy, dist, relief, slope	6
21	Model 21: lon, lat, bathy, dist, relief	lon, lat, bathy, dist, relief	5
22	Model 22: lon, lat, bathy, dist	lon, lat, bathy, dist	4
23	Model 23: lon, lat, dist	lon, lat, dist	3

Application to Marine Environment

Effects of input variables

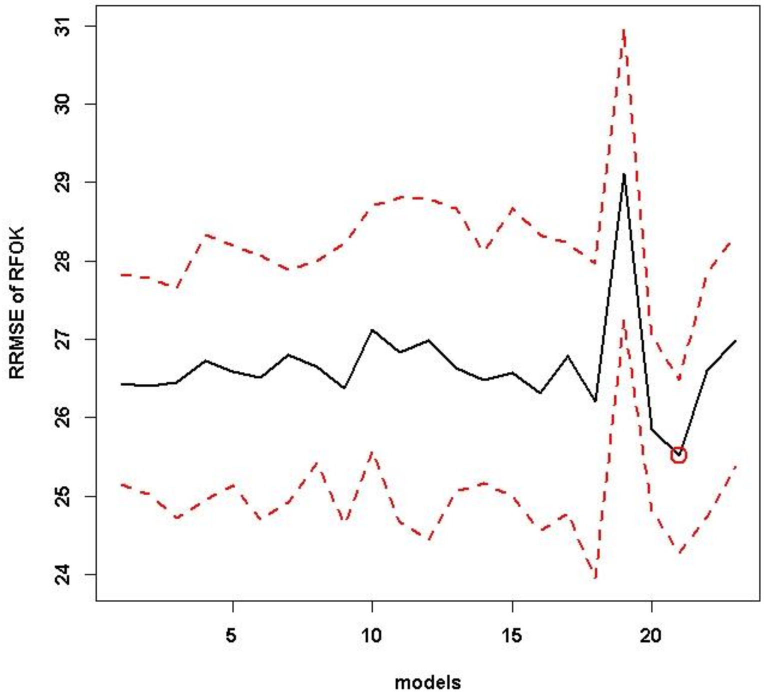
Gravel content: 22 models

	Modelling.process	Predictors	No.predictors
1	Model 1: All 21 predictors	All 21 variables	21
2	Model 2: - sslope from model 1	lon, lat, bathy, dist, relief, slope, sbathy, cbathy, sdist.coast, cdist.coast, srelief, crelief, cslope, slat, clat, slon, clon, latlon, latslon, slatlon	20
3	Model 3: - cslope from model 2	lon, lat, bathy, dist, relief, slope, sbathy, cbathy, sdist.coast, cdist.coast, srelief, crelief, slat, clat, slon, clon, latlon, latslon, slatlon	19
4	Model 4: - clat from model 3	lon, lat, bathy, dist, relief, slope, sbathy, cbathy, sdist.coast, cdist.coast, srelief, crelief, slat, slon, clon, latlon, latslon, slatlon	18
5	Model 5: - relief and crelief from model 4	lon, lat, bathy, dist, slope, sbathy, cbathy, sdist.coast, cdist.coast, srelief, slat, slon, clon, latlon, latslon, slatlon	16
6	Modle 6: - latlon and slatlon from model 5	lon, lat, bathy, dist, slope, sbathy, cbathy, sdist.coast, cdist.coast, srelief, slat, slon, clon, latslon	14
7	Model 7: - slope from model 6	lon, lat, bathy, dist, sbathy, cbathy, sdist.coast, cdist.coast, srelief, slat, slon, clon, latslon	13
8	Model 8: - cdist.coast from model 7	lon, lat, bathy, dist, sbathy, cbathy, sdist.coast, srelief, slat, slon, clon, latslon	12
9	Model 9: - latslon from model 8	lon, lat, bathy, dist, sbathy, cbathy, sdist.coast, srelief, slat, slon, clon	11
10	Modle 10: - cbathy from model 9	lon, lat, bathy, dist, sbathy, sdist.coast, srelief, slat, slon, clon	10
11	Model 11: - slat from model 10	lon, lat, bathy, dist, sbathy, sdist.coast, srelief, slon, clon	9
12	Model 12: - lat from model 11	lon, bathy, dist, sbathy, sdist.coast, srelief, slon, clon	8
13	Model 13: - srelief from model 12	lon, bathy, dist, sbathy, sdist.coast, slon, clon	7
14	Model 14: - sbathy from model 13	lon, bathy, dist, sdist.coast, slon, clon	6
15	Model 15: - clon from model 14	lon, bathy, dist, sdist.coast, slon	5
16	Modle 16: - slon from model 15	lon, bathy, dist, sdist.coast	4
17	Model 17: - sdist.coast from model 16	lon, bathy, dist	3
18	Model 18: - bathy from model 17	lon, dist	2
19	Model 19: - lon from model 18	dist	1
20	Model 20: lon, lat, bathy, dist, relief, slope	lon, lat, bathy, dist, relief, slope	6
21	Model 21: lon, lat, bathy, dist, slope	lon, lat, bathy, dist, slope	5
22	Model 22: lon, lat, bathy, dist	lon, lat, bathy, dist	4

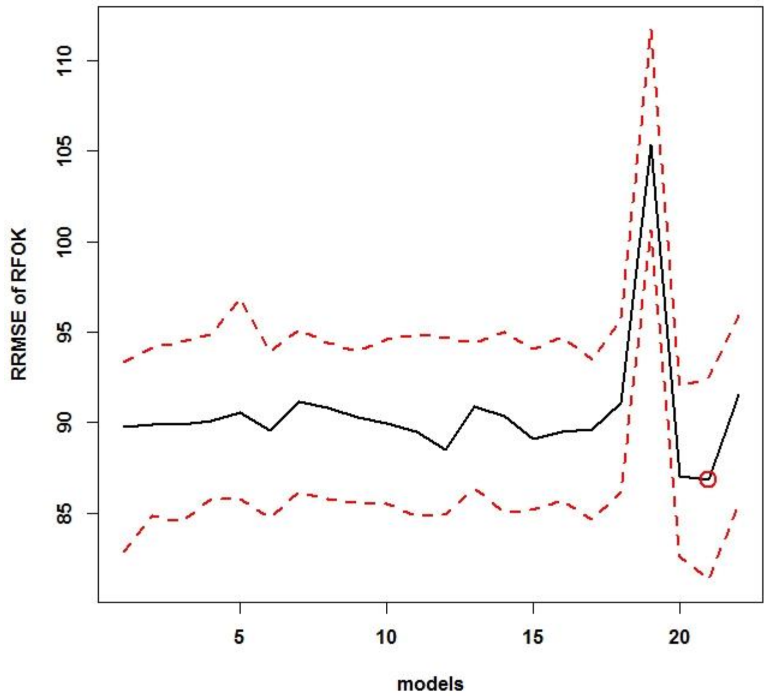
Application to Marine Environment

Effects of input variables

Sand content: 23 models



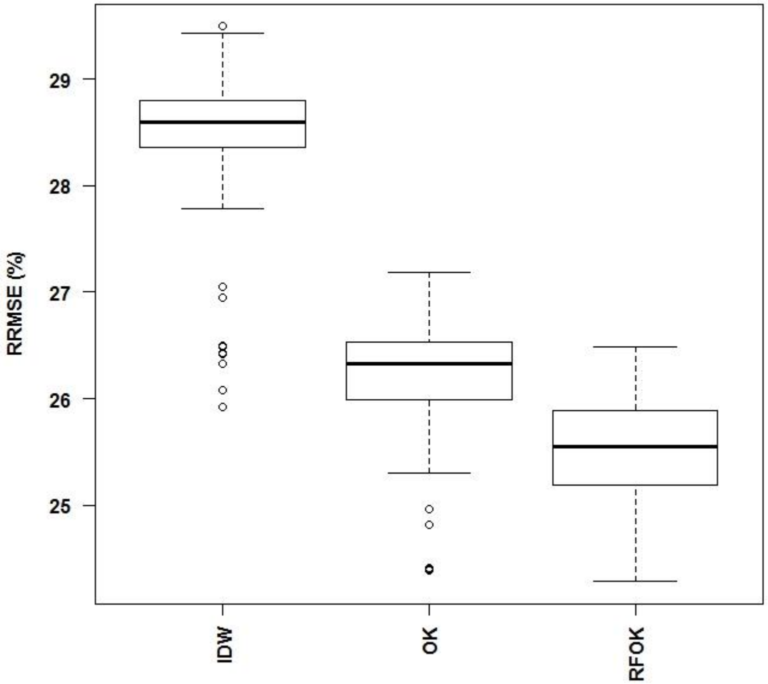
Gravel content: 22 models



Application to Marine Environment

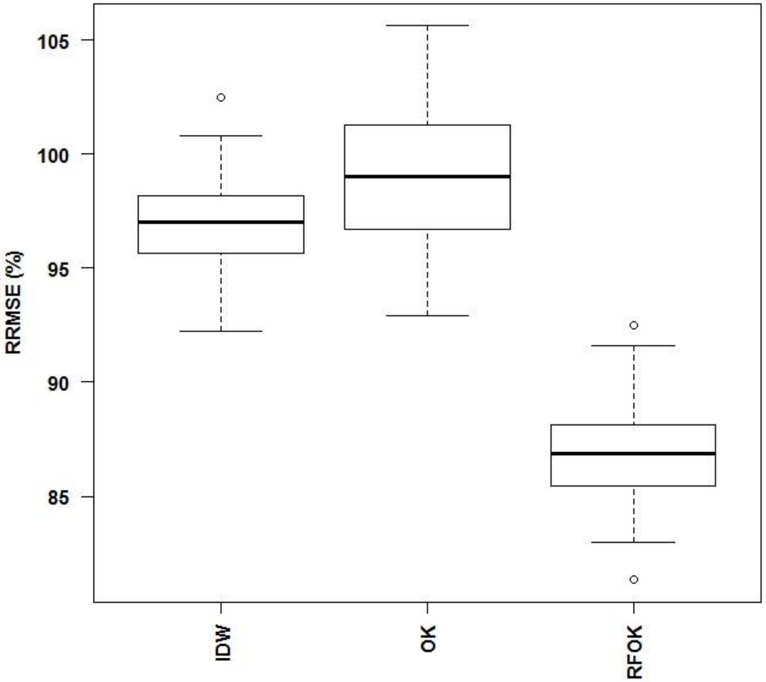
Effects of Methods

Sand content



RRPE: 10.2%

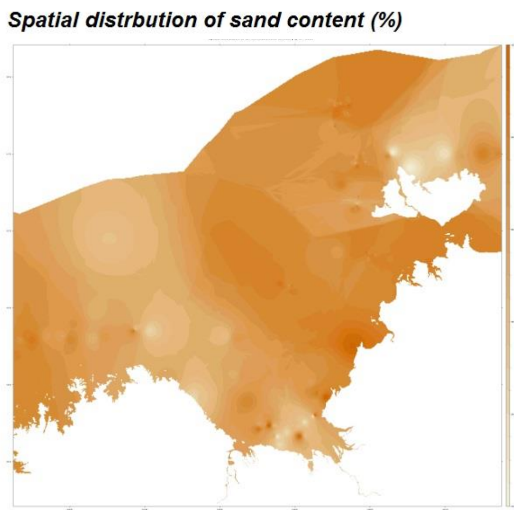
Gravel content

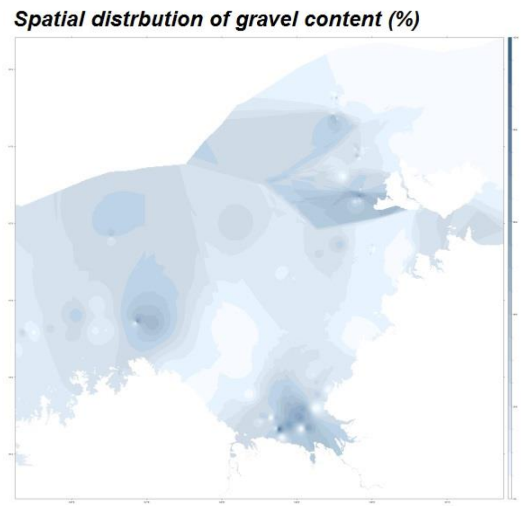


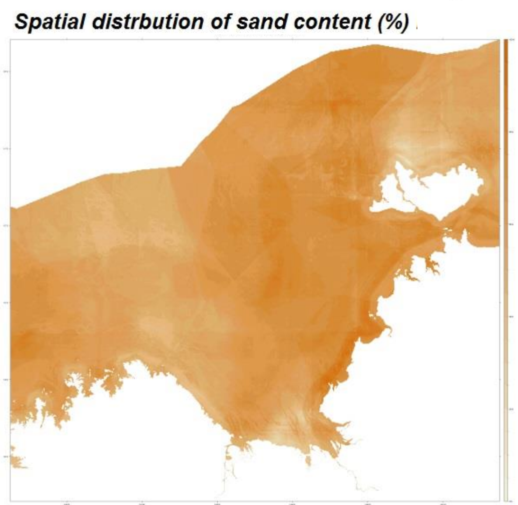
RRPE: 10.3%

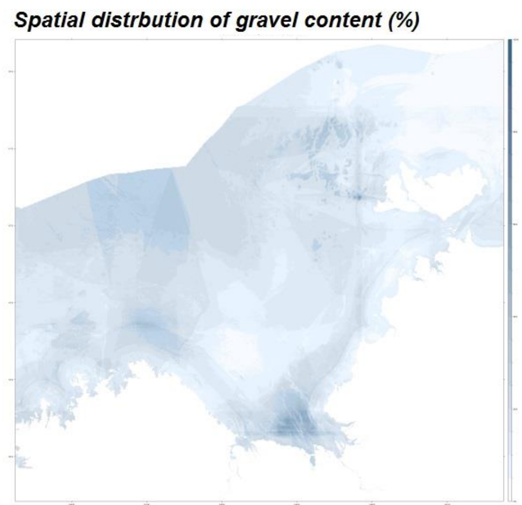
Application to Marine Environment

Spatial predictions of IDW and RFOK







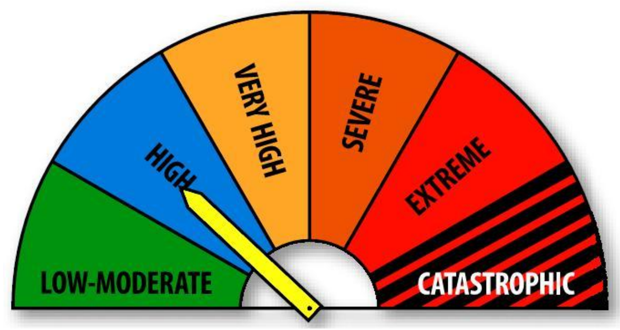


Application to the terrestrial environment

Fire Weather Danger

One of the most commonly used Fire Weather Danger indicator in Australia is the McArthur Forest Fire Danger Index (FFDI).

Category	Forest Fire Danger Index
Catastrophic (Code Red)	100 +
Extreme	75 – 99
Severe	50 – 74
Very high	25 - 49
High	12 – 24
Low to moderate	0 - 11



Fire Danger Rating

Application to the terrestrial environment

Quantifying natural hazards

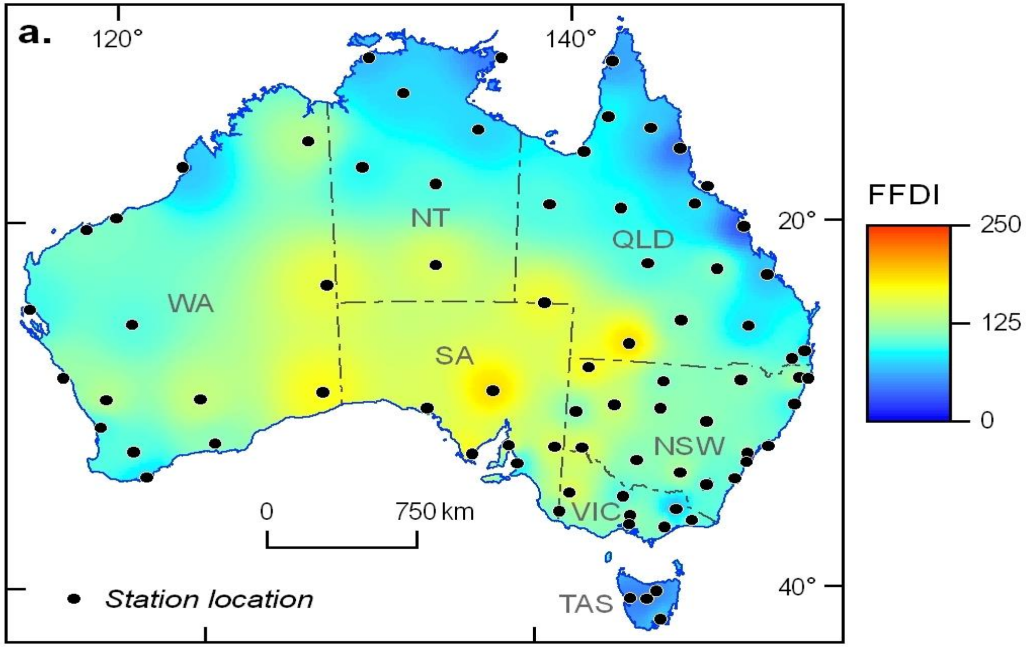
Average Recurrence Interval (Return period).

If a given value (return level) of some natural phenomenon such as wind speed, temperature or precipitation is exceeded with probability 'p' on average once a year, the Return Period (RP) corresponding to this value is 1/p years.

Example. The average annual probability of exceeding a gust wind speed of 45 m/s at Sydney Airport is 0.002, we can say that the 500-year RP (1/0.002) of gust wind speed at this location is 45 m/s, i.e. it is expected that the value 45 m/s is exceeded at Sydney Airport, on average, once every 500 years.

Application to the terrestrial environment

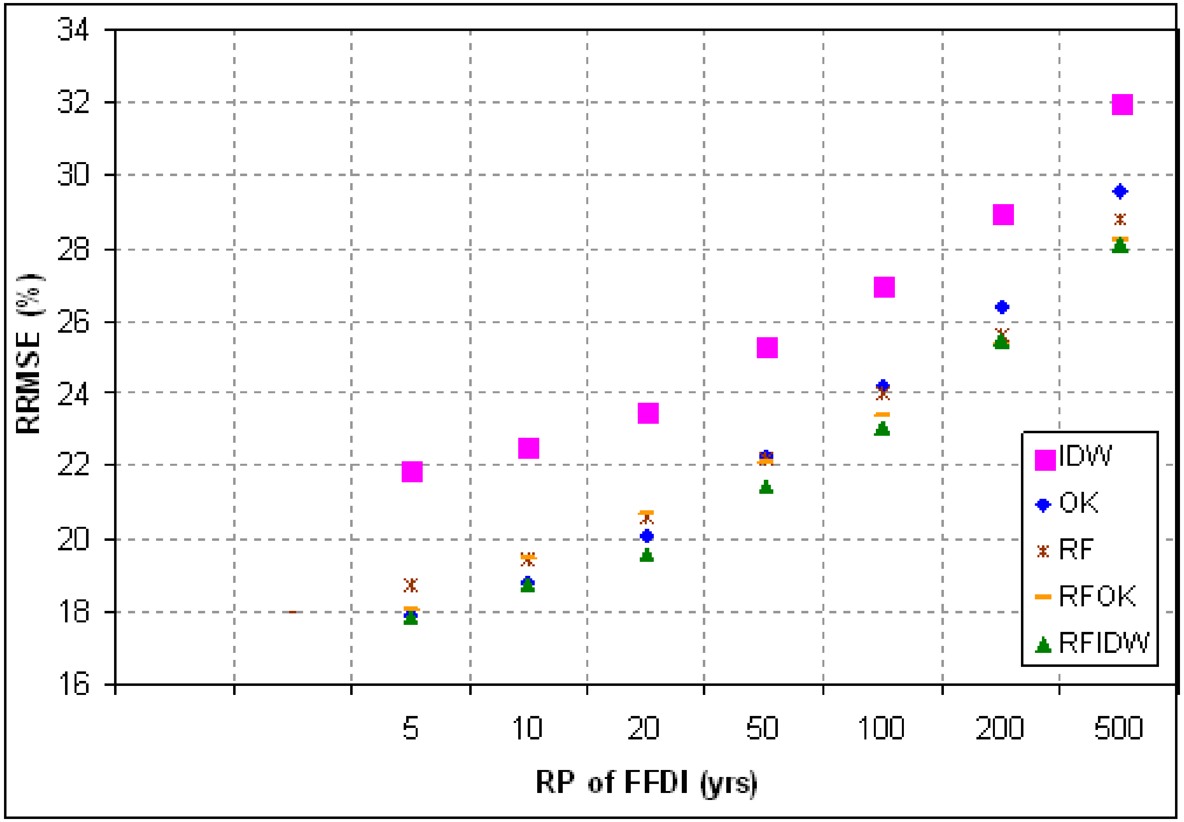
Samples of FFDI (*n*=78)



Application to the terrestrial environment

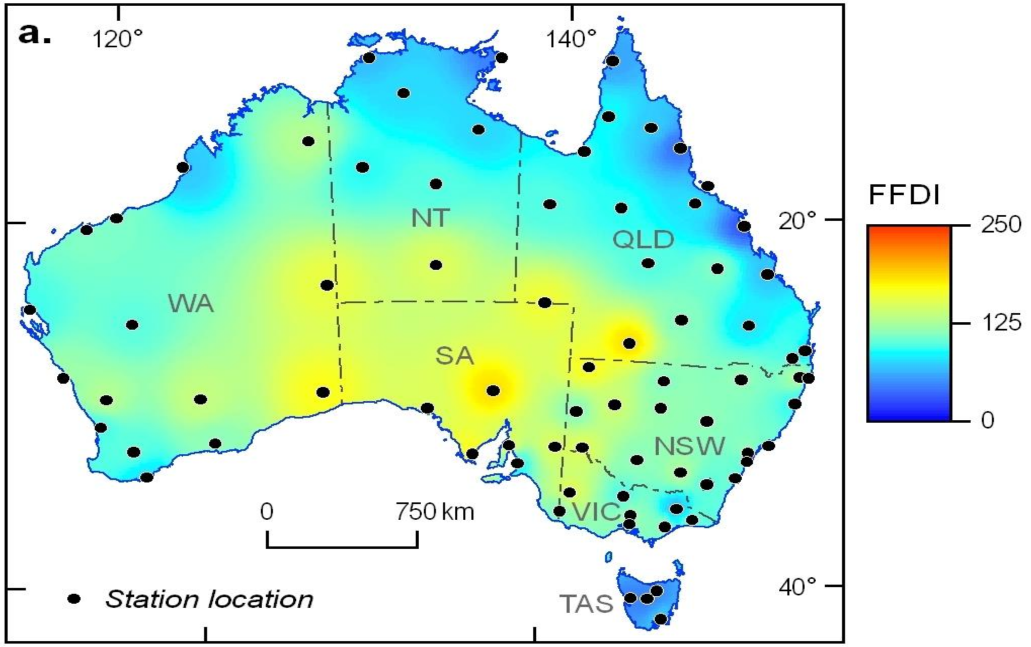
Variable	Name
Annual mean temperature	T_mean
Summer mean temperature	T_mean_djf
Autumn mean temperature	T_mean_mam
Winter mean temperature	T_mean_jja
Spring mean temperature	T_mean_son
Annual maximum temperature	T_max
Summer maximum temperature	T_max_djf
Autumn maximum temperature	T_max_mam
Winter maximum temperature	T_max_jja
Spring maximum temperature	T_max_son
Annual minimum temperature	T_min
Summer minimum temperature	T_min_djf
Autumn minimum temperature	T_min_mam
Winter minimum temperature	T_min_jja
Spring minimum temperature	T_min_son
Annual mean precipitation	Rain_mean
Summer mean precipitation	Rain_djf
Autumn mean precipitation	Rain_mam
Winter mean precipitation	Rain_jja
Annual mean relative humidity	RH_mean
Summer mean relative humidity	RH_djf
Autumn mean relative humidity	RH_mam
Winter mean relative humidity	RH_jja
Spring mean relative humidity	RH_son
Annual mean pan evaporation	Evp_mean
Summer mean pan evaporation	Evp_djf
Autumn mean pan evaporation	Evp_mam
Winter mean pan evaporation	Evp_jja
Spring mean pan evaporation	Evp_son
Mean enhanced vegetation index	EVI_mean
Maximum enhanced vegetation index	EVI_max
Minimum enhanced vegetation index	EVI_min
Annual mean wind speed	Wind
Elevation (above sea level)	Elevation
Latitude/Longitude	Lat/Lon

Application to the terrestrial environment



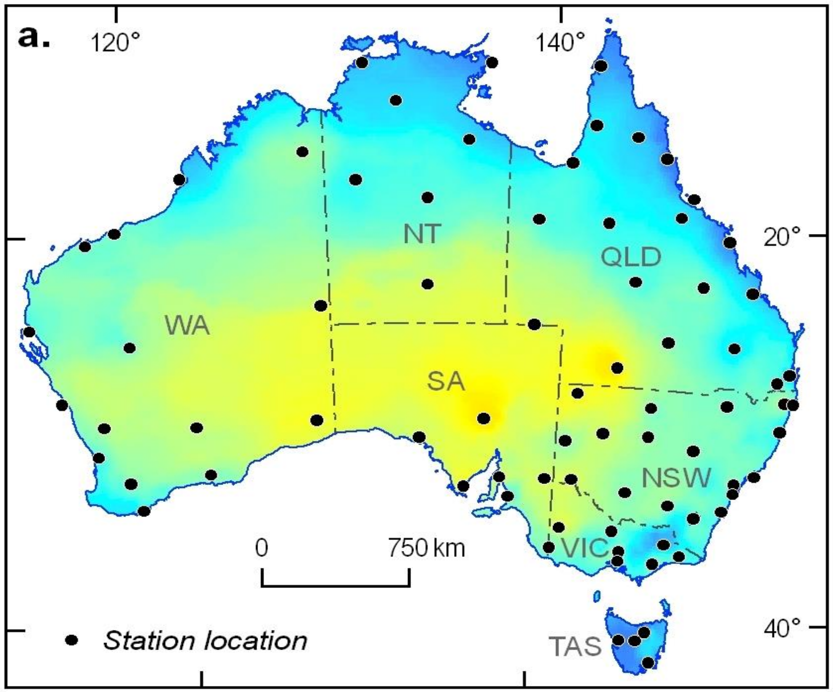
RRMSE (%) based on leave-one-out cross-validation

Application to the terrestrial environment

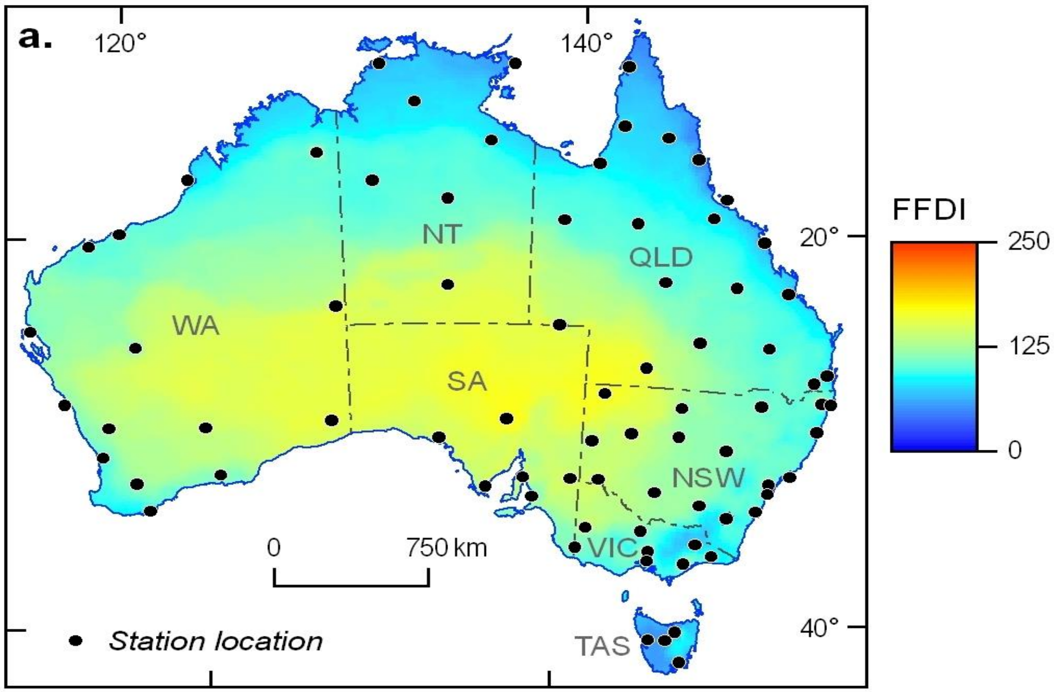


Spatial predictions of the 50-yr RP of FFDI using IDW

Application to the terrestrial environment



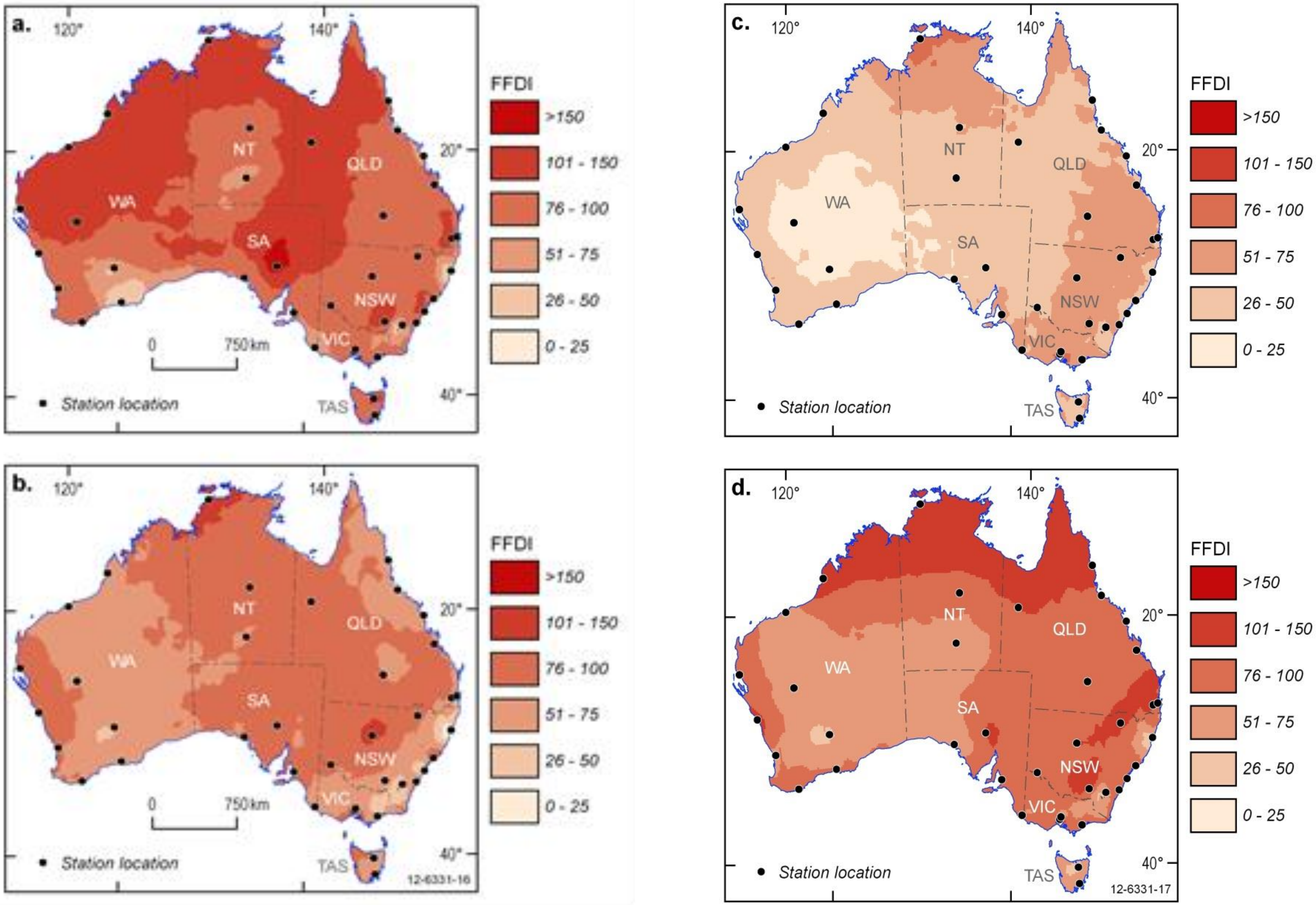
RF-IDW



RF-OK

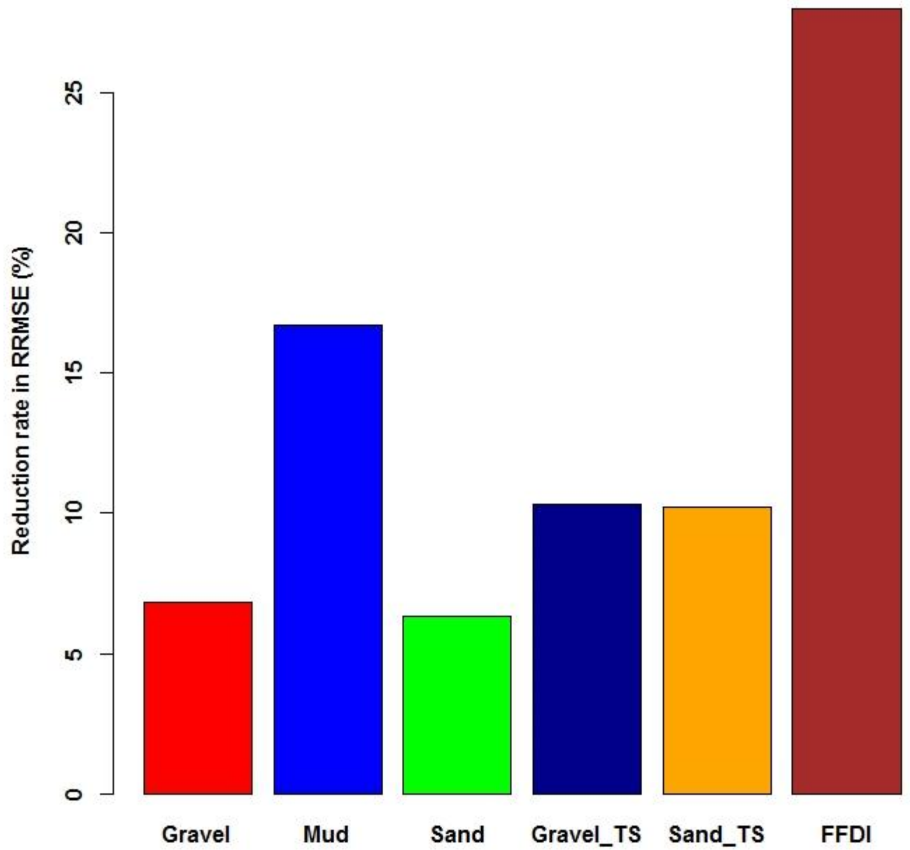
Spatial predictions of the 50-yr RP of FFDI

Application to the terrestrial environment



Predictions of the 50-yr RP of FFDI. a) Summer. b) Autumn. c) Winter. d) Spring

RRPE (%) for spatial predictions of seabed sediment in the previous studies (Li et al. 2010, 2011a, b, c & 2012) and current study (Li 2013a), and of FFDI (Sanabria et al. 2013)



1) These hybrid methods seem not data specific, but their models are. Therefore, best model should be developed according to individual situation.
2) Model selection is required for RF and the hybrid method in order to find an optimal predictive model.
3) The most accurate predictions were obtained using RFOK and RFIDW, with a RRPE of 10% for seabed sediment and 28% for FFDI when compared to IDW.
4) These methods have been applied to about 20 datasets in marine and terrestrial environments with promising results. They are recommended not only for environmental sciences but also for other disciplines.
5) The development of the hybrid methods has opened an alternative source of methods for spatial prediction.
6) More machine learning methods are expected to be introduced to and new hybrid methods are expected to be developed for and applied to spatial predictive modelling in the future.

References

Li, J. and A. Heap (2008). A Review of Spatial Interpolation Methods for Environmental Scientists, Geoscience Australia. Record 2008/23, 137pp.

Li J., Heap A., 2011. A review of comparative studies of spatial interpolation methods: performance and impact factors. *Ecological Informatics* 6: 228-241.

Li J., Heap A., Potter A., Daniell J.J., 2011a. Predicting Seabed Mud Content across the Australian Margin II: Performance of Machine Learning Methods and Their Combination with Ordinary Kriging and Inverse Distance Squared. Geoscience Australia: Record 2011/07 69.

Li J., Heap A.D., Potter A., Daniell J., 2011b. Application of machine learning methods to spatial interpolation of environmental variables. *Environmental Modelling & Software* 26: 1647-1659.

Li J., Heap A.D., Potter A., Huang Z., Daniell J., 2011c. Can we improve the spatial predictions of seabed sediments? A case study of spatial interpolation of mud content across the southwest Australian margin. *Continental Shelf Research* 31: 1365-1376.

Li J., Potter A., Huang Z., Daniell J.J., Heap A., 2010. Predicting Seabed Mud Content across the Australian Margin: Comparison of Statistical and Mathematical Techniques Using a Simulation Experiment. Geoscience Australia: Record 2010/11 146.

Li, J., Potter, A., Huang, Z. and Heap, A. D., 2012. Predicting Seabed Sand Content across the Australian Margin Using Machine Learning and Geostatistical Methods. Geoscience Australia, Record 2012/48, 115 pp.

Li J (2013a) Predicting the spatial distribution of seabed gravel content using random forest, spatial interpolation methods and their hybrid methods. *The International Congress on Modelling and Simulation (MODSIM) 2013*. Adelaide.

Li J (2013b) Predictive Modelling Using Random Forest and Its Hybrid Methods with Geostatistical Techniques in Marine Environmental Geosciences. In: Christen P, Kennedy P, Liu L, Ong K-L, Stranieri A et al., edit. *The proceedings of the Eleventh Australasian Data Mining Conference (AusDM 2013*), Canberra, Australia, 13-15 November 2013: Conferences in Research and Practice in Information Technology, Vol. 146.

Li J, Heap AD (2014) Spatial interpolation methods applied in the environmental sciences: A review. *Environmental Modelling & Software* 53: 173-189.

Okun, O., Priisalu, H., 2007. Random forest for gene expression based cancer classification: overlooked issues. In: Martí, J., Benedí, J.M., Mendonça, A.M., Serrat, J. (Eds.), *Pattern Recognition and Image Analysis: Third Iberian Conference*. IbPRIA 2007 Lecture Notes in Computer Science, Girona, Spain, pp. 4478, 4483-4490.

Sanabria LA, Qin X, Li J, Cechet RP, Lucas C (2013) Spatial interpolation of McArthur's forest fire danger index across Australia: observational study. *Environmental Modelling & Software* 50: 37-50.

Thank you!

Phone: +61 2 6249 9111
Web: www.ga.gov.au
Email: feedback@ga.gov.au
Address: Cnr Jerrabomberra Avenue and Hindmarsh Drive, Symonston ACT 2609
Postal Address: GPO Box 378, Canberra ACT 2601