Rainfall forecast model using Support Vector Machine (SVM) for extreme monsoon rainfall conditions in an urban area: Mumbai, India

By
Vinay S Nikam
Kapil Gupta
IIT Bombay

9th Urban Drainage Modelling

3-7 September, 2012, Belgrade, Serbia

IIT Bombay

Outline

1. Aims and objectives
2. Rainfall forecasting with support vector machine (SVM)
3. Results and discussions

Objective

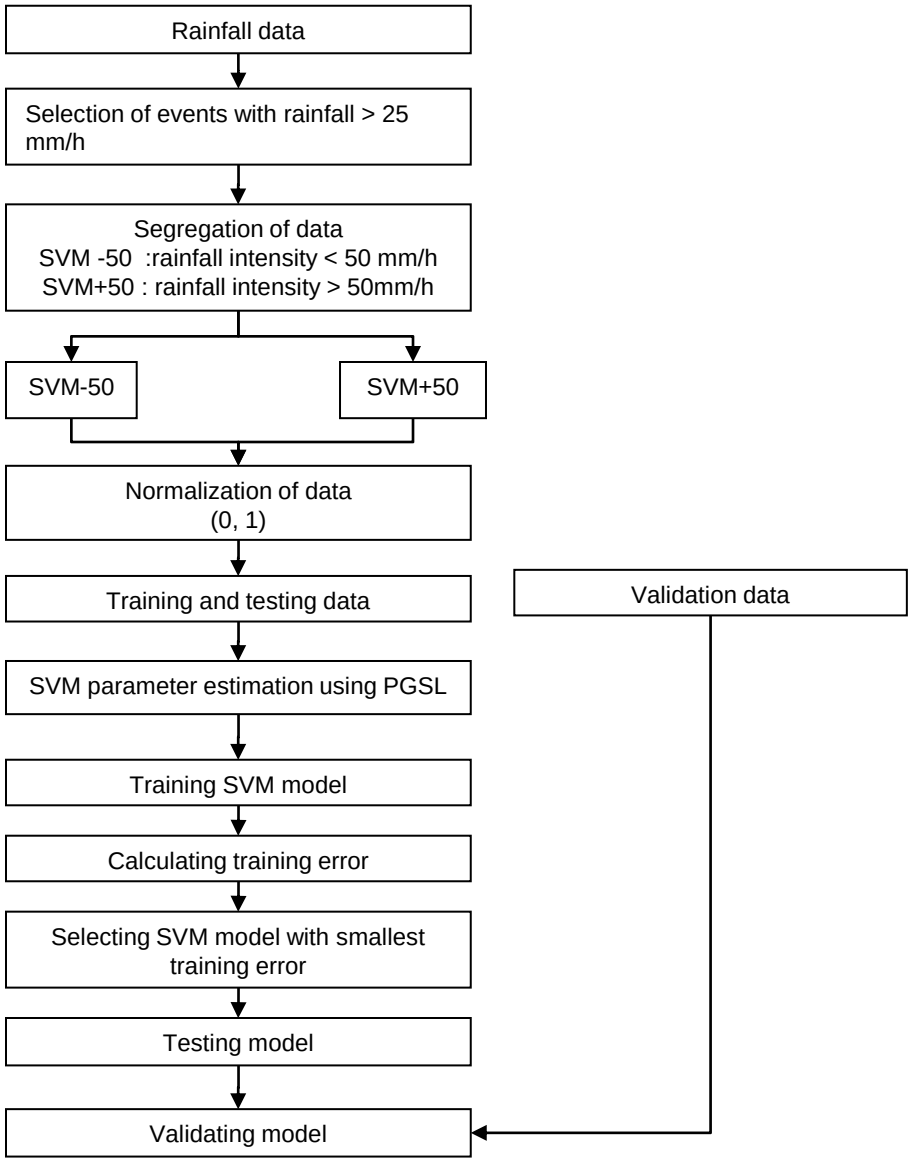
1. To develop a model for short term (15-30 minutes) forecasting of high intensity rainfall using SVM

- Population – 11.91 Million (2001 census)
- Annual average rainfall 2300 mm (Santacruz)
- 30% to 35 % rainfall occurs in 2-3 events
- Rainfall data collected daily by 2 Symond's rain gauges before 26th July 2005 flood
- 35 rain gauges installed in June 2006 with an average rain gauge intensity of 1 per 16 km^2

- Symond's rain gauge
- Tipping bucket rain gauge

IIT Bombay

Flow chart for rainfall forecast model



PGSL: Probabilistic Global Search Lausanne

IIT Bombay

Training, Testing and Validation Data

S. No.	Training data			Testing data		
	No. of years	Year	No. of events	No. of years	Year	No. of events
1	3	2007, 2008, 2009	34	1	2010	26

Model has been validated for 4 rainfall events of the year 2011 (31-07-11, 27-08-11, 28-08-11, 29-08-11)

Data resolution

1. If the objective is to forecast 5-min periods into the future, then the best data resolution to be used is 5-min. Also higher the resolution of the data, the higher the prediction error (Abdulhai et al., 2002)
2. Resolution of the data should be equal to that of the data to be predicted (Clark, 2003; Abdulhai et al., 2002; Chen and Grant-Muller, 2001)

Values of parameters used for different lead times

S. No.	Lead time	SVM		SVM-50		SVM+50	
		γ	σ	γ	σ	γ	σ
1	5-min	0.4075	0.3174	0.1807	0.3631	0.5590	0.7976
2	10-min	0.2313	0.3523	0.3416	0.6562	0.7326	0.3723
3	15-min	0.2070	0.3430	0.4417	0.7318	0.3118	0.0978
4	20-min	0.2730	0.2730	0.4925	0.4925	0.2627	0.2627
5	25-min	0.2730	0.2730	0.7677	0.3370	0.3203	0.1406
6	30-min	0.3330	0.0660	0.8235	0.1738	0.3210	0.0678

γ regularization parameter
σ width of radial basis kernel function

The parameter γ controls the trade off between errors of the SVM on training data and margin maximization

Performance indicators

$$Correlation\ coefficient = \frac{\sum_{i=1}^{n}(a_i - \bar{a}) * (f_i - \bar{f})}{\sqrt{\sum_{i=1}^{n}(a_i - \bar{a})^2 * \sum_{i=1}^{n}(f_i - \bar{f})^2}}$$

$$Root\ mean\ square\ error\ (RMSE) = \sqrt{\frac{\sum_{i=1}^{n}(a_i - f_i)^2}{n}}$$

$$Nash\ Sutcliffe\ efficiency\ coefficient = 1 - \frac{\sum_{i}^{n}(a_i - f_i)^2}{\sum_{i}^{n}(a_i - \bar{a})^2}$$

$$Index\ of\ agreement\ (IA) = 1 - \frac{\sum_{i=1}^{n}(f_i - a_i)^2}{\sum_{i=1}^{n}(|f_{i-} \bar{a}| + |a_{i-} \bar{a}|)^2}$$

$$Coefficient\ of\ determination = \left(\frac{\sum_{i=1}^{n}(a_i - \bar{a})(f_i - \bar{f})}{\left[\sum_{i=1}^{n}(a_i - \bar{a})^2 \sum_{i=1}^{n}(f_i - \bar{f})^2\right]^{0.5}} \right)^2$$

$$Relative\ error\ in\ peak = \frac{\max f_i - \max a_i}{\max f_i} \times 100$$

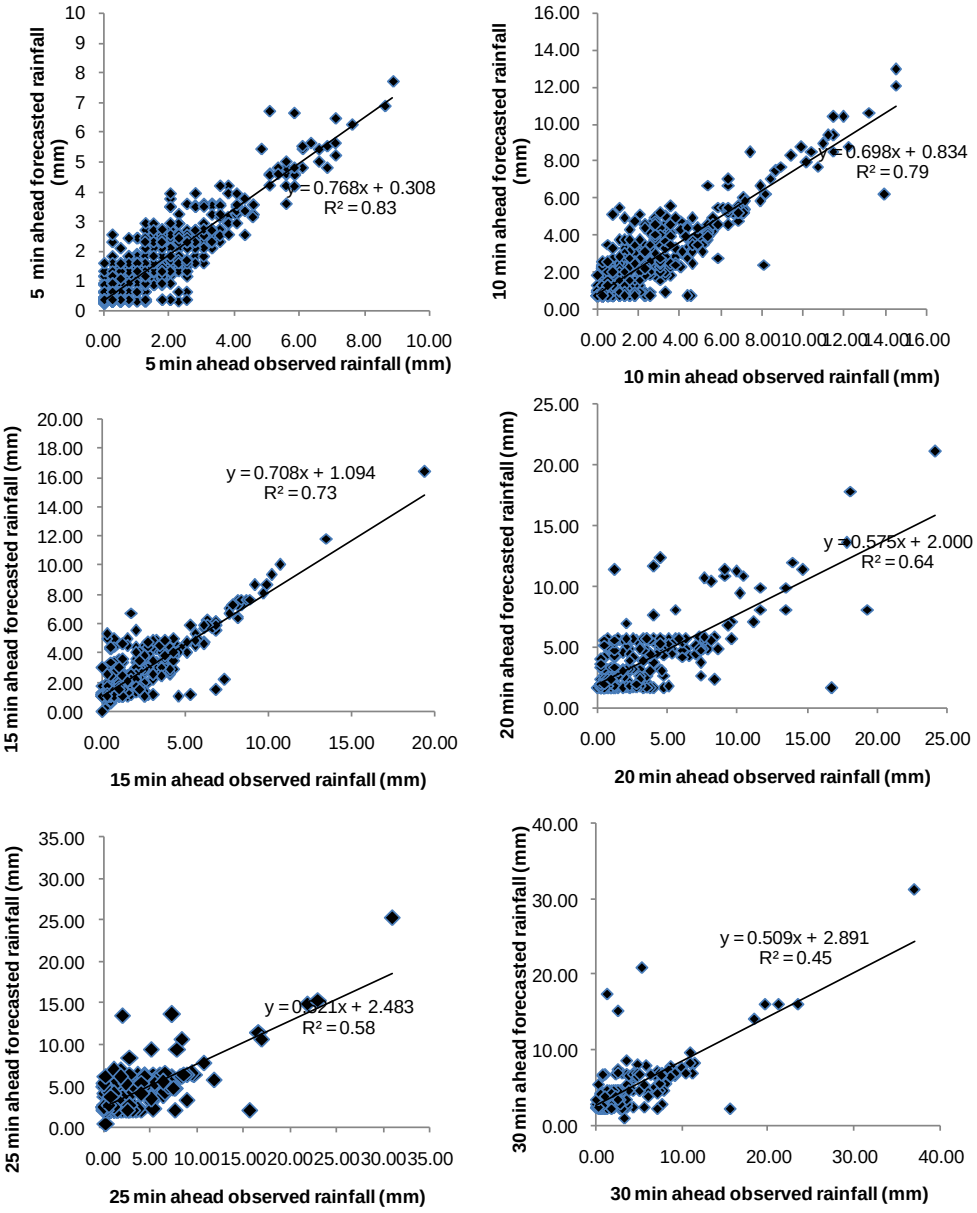
where,

a_i and f_i represent the actual and forecast mean rainfall, $\bar{a}$ and $\bar{f}$ represent the actual and forecast mean rainfall

Performance statistics for different models and lead times – Testing

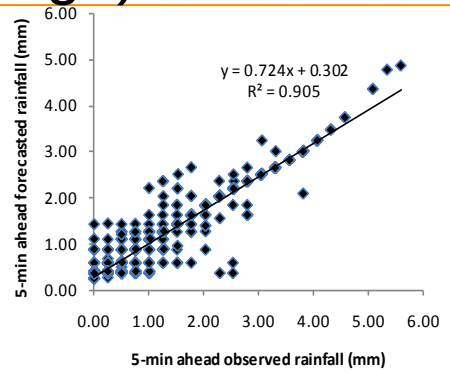
Lead time	CC	RMSE	Nash (CE)	IA	CoD	RE
	SVM					
5-min	0.9170	0.53	0.8259	0.9471	0.8375	0.1316
10-min	0.8956	1.120	0.7702	0.9241	0.7972	0.1075
15-min	0.8526	1.59	0.6992	0.9003	0.7374	0.1489
20-min	0.8057	2.83	0.5290	0.8636	0.6458	0.1916
25-min	0.7609	2.52	0.5123	0.8148	0.5801	0.2341
30-min	0.6760	3.37	0.3733	0.7812	0.4507	0.3146
	SVM-50					
5-min	0.9514	0.36	0.8624	0.9551	0.9051	0.1252
10-min	0.9250	0.82	0.7953	0.9262	0.8588	0.1075
15-min	0.9228	1.18	0.7522	0.9166	0.8516	0.1491
20-min	0.9071	1.51	0.7665	0.9066	0.8229	0.1557
25-min	0.8890	1.83	0.7495	0.9012	0.7904	0.1859
30-min	0.8867	2.06	0.7218	0.8982	0.7862	0.2153
	SVM+50					
5-min	0.9626	0.51	0.9106	0.9731	0.9266	0.1002
10-min	0.9603	0.97	0.8976	0.9685	0.9229	0.1121
15-min	0.9498	2.10	0.8313	0.9431	0.9012	0.1296
20-min	0.9356	2.38	0.8142	0.9306	0.8754	0.1383
25-min	0.9068	2.60	0.7408	0.8992	0.8222	0.4254
30-min	0.8698	2.62	0.7101	0.8797	0.7565	0.4736

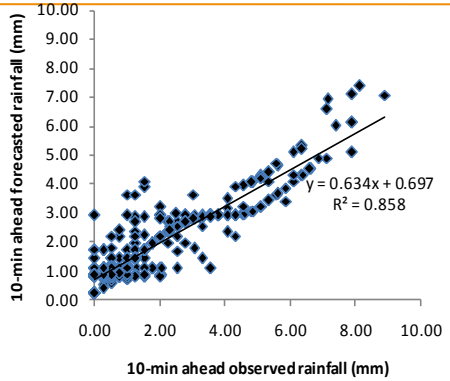
Scatter plot of observed and forecasted rainfall for SVM model (testing stage)

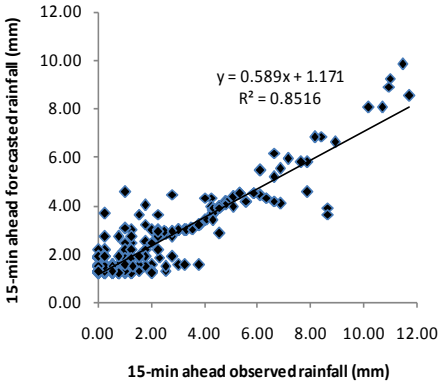


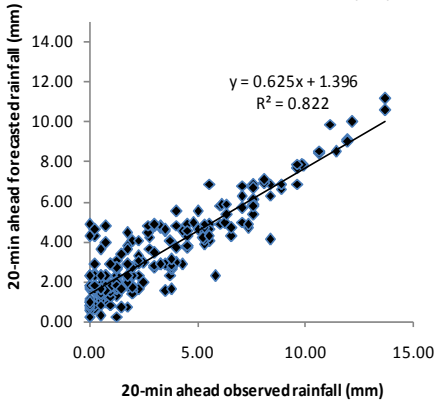
IIT Bombay

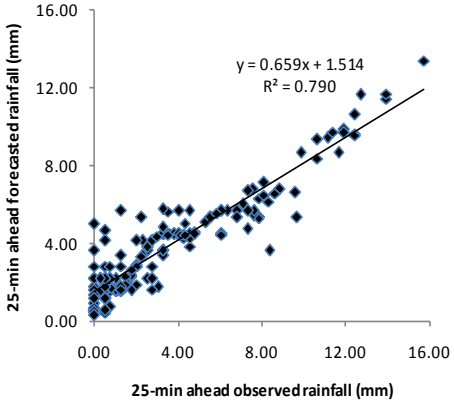
Scatter plot of observed and forecasted rainfall for SVM-50 model (testing stage)

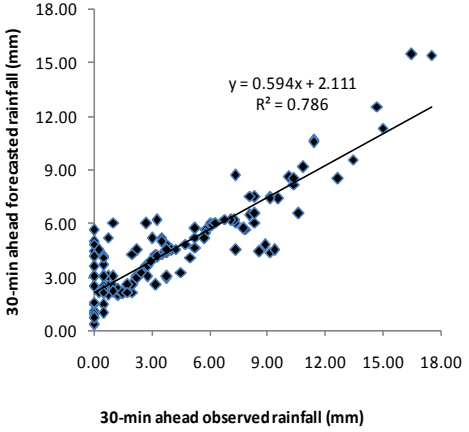






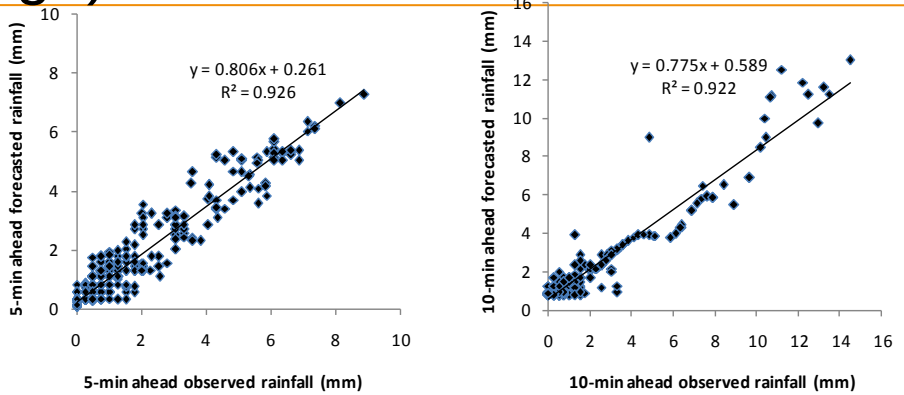


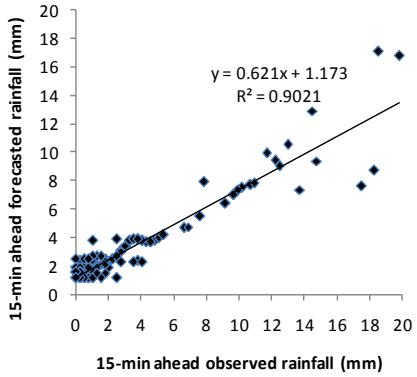


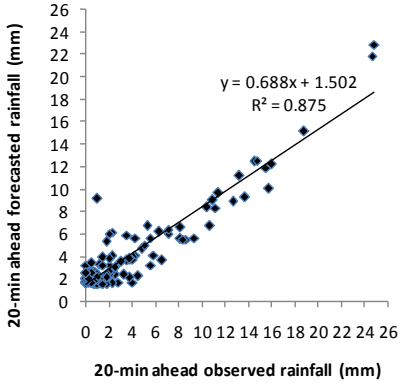


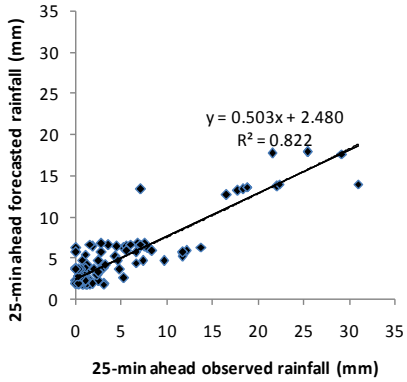
IIT Bombay

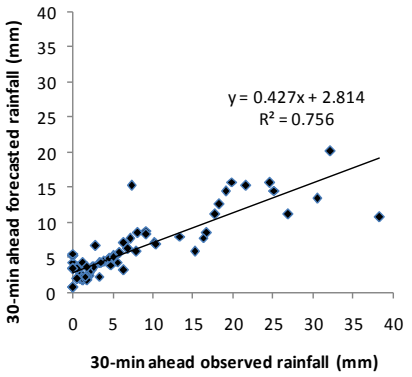
Scatter plot of observed and forecasted rainfall for SVM+50 model (testing stage)









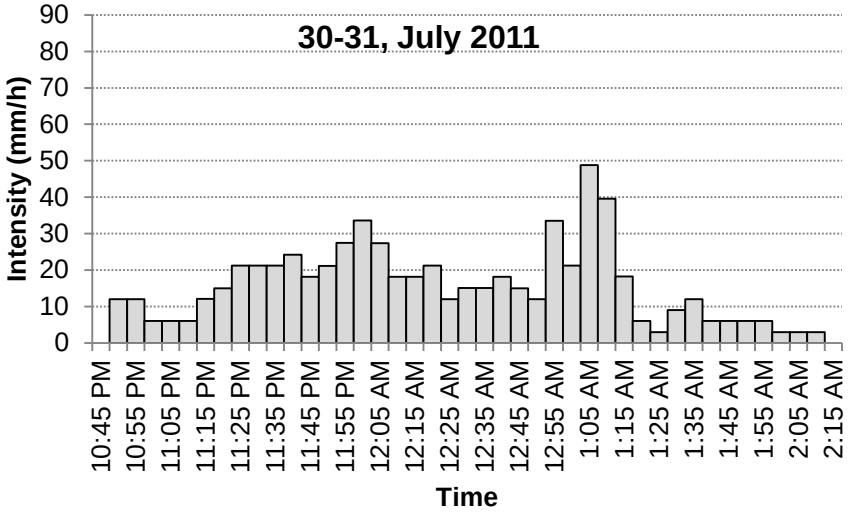


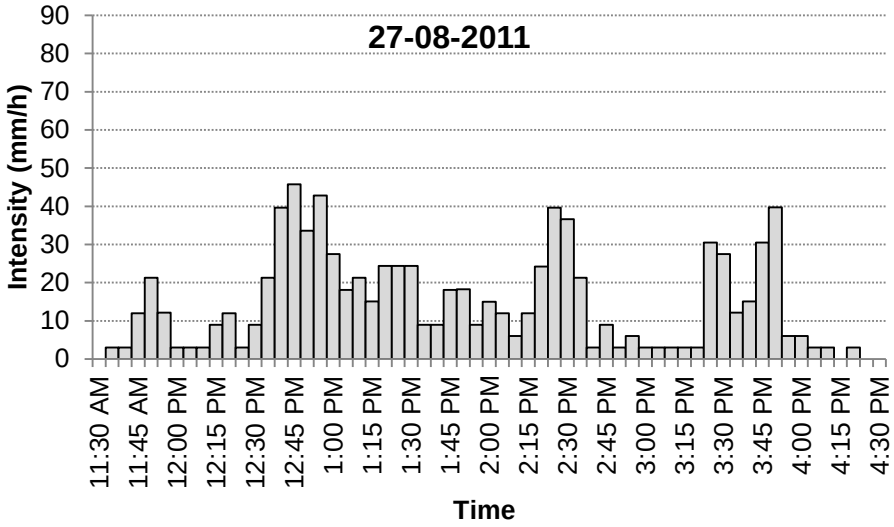
IIT Bombay

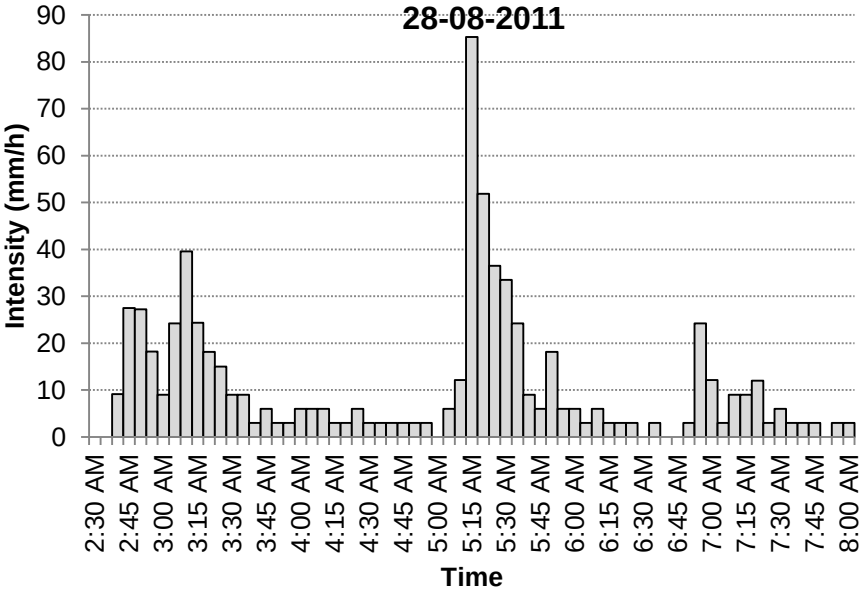
Rainfall events used for validation of forecast model

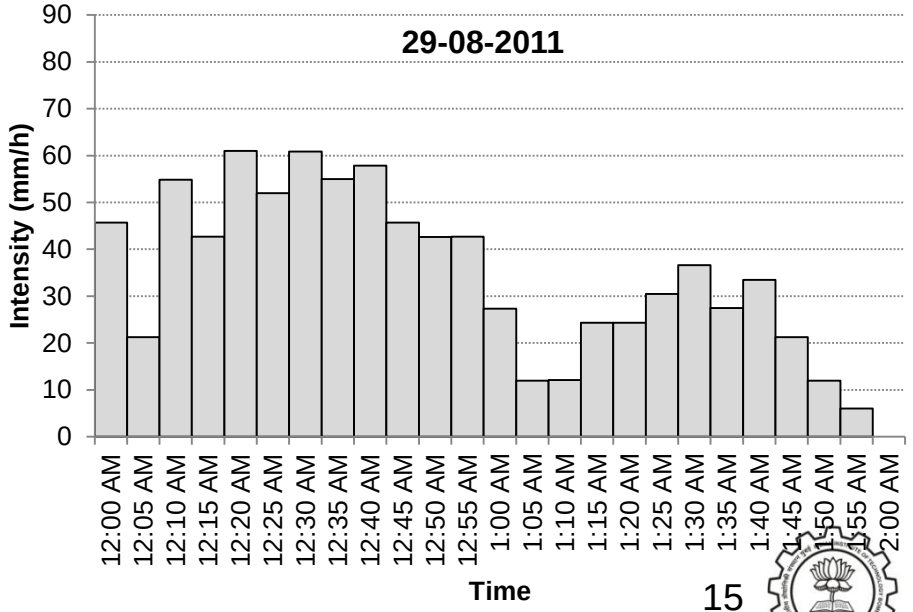
S. No.	Date	Time	Rainfall duration (h:min)	Total event rainfall (mm)
1	30-7-2011 and 31-7-2011	10:45 PM to 2:15 AM	3:30	54.50
2	27-08-2011	11:30 AM to 4:30 PM	5:00	72.75
3	28-08-2011	2:30 AM to 8:00 AM	5:30	58.75
4	29-08-2011	0:00 AM to 2:00 AM	2:00	70.75

Hyetograph recorded at 5-min time-step for the events used for validating forecast model









IIT Bombay

Performance statistics of 5-30 min ahead rainfall forecast for July 30-31, 2011 and August 27, 2011 rainfall event

July 30-31, 2011

	CC	RMSE	Nash (CE)	IoA	CoD	RE
5-min	0.9591	0.28	0.8940	0.9666	0.9197	0.1337
10-min	0.9588	0.65	0.8843	0.9628	0.9192	0.1451
15-min	0.9402	0.68	0.8841	0.9591	0.9015	0.1144
20-min	0.9358	1.23	0.8984	0.9301	0.8947	0.1861
25-min	0.8963	2.40	0.7577	0.7992	0.8043	0.3159
30-min	0.8861	3.25	0.7454	0.7751	0.7966	0.3815

August 27, 2011

	CC	RMSE	Nash (CE)	IoA	CoD	RE
5-min	0.9729	0.29	0.9676	0.9551	0.9481	0.1049
10-min	0.9513	0.63	0.8959	0.9579	0.9051	0.1152
15-min	0.9470	1.29	0.8729	0.9355	0.8968	0.1457
20-min	0.9616	1.56	0.6324	0.9791	0.9240	0.2114
25-min	0.8310	2.33	0.5717	0.8013	0.6906	0.2674
30-min	0.9244	2.70	0.6067	0.8593	0.8563	0.3323

Performance statistics of 5-30 min ahead rainfall forecast for August 28, 2011 and August 29, 2011 rainfall event

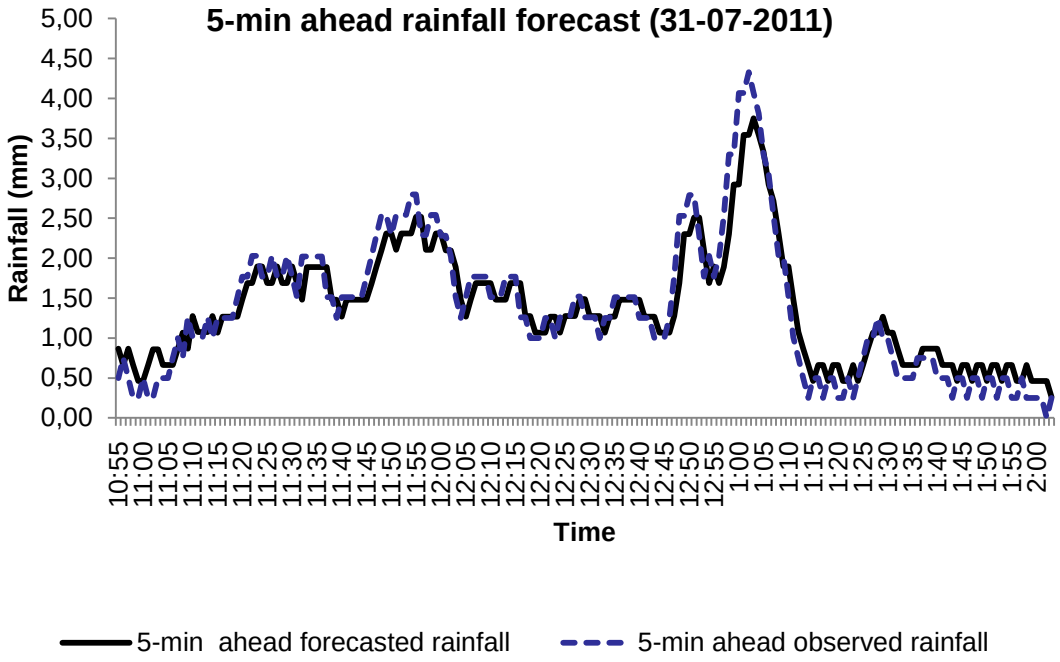
August 28, 2011

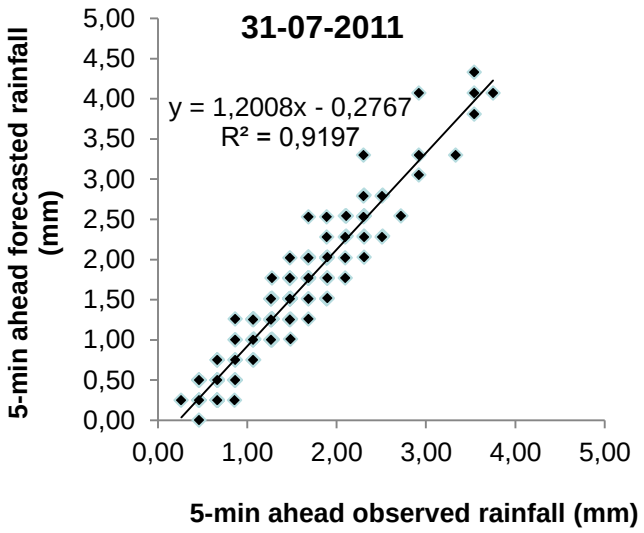
	CC	RMSE	Nash (CE)	IoA	CoD	RE
5-min	0.9065	1.00	0.9332	0.9800	0.9609	0.1153
10-min	0.9528	0.74	0.8693	0.9585	0.9097	0.1038
15-min	0.9895	0.96	0.9028	0.9697	0.9728	0.1768
20-min	0.7780	2.16	0.8090	0.8411	0.6050	0.1908
25-min	0.7627	2.14	0.6826	0.8860	0.7705	0.2069
30-min	0.7627	3.07	0.7183	0.6826	0.5936	0.3775

August 29, 2011

	CC	RMSE	Nash (CE)	IoA	CoD	RE
5-min	0.9591	0.57	0.9530	0.8435	0.9505	0.1292
10-min	0.9588	1.31	0.9245	0.7378	0.9127	0.1460
15-min	0.9402	2.16	0.9427	0.5978	0.8605	0.2201
20-min	0.9358	2.58	0.9197	0.7161	0.8980	0.2293
25-min	0.8963	5.05	0.7578	0.4964	0.6647	0.3871
30-min	0.8862	6.82	0.7935	0.3924	0.6542	0.4335

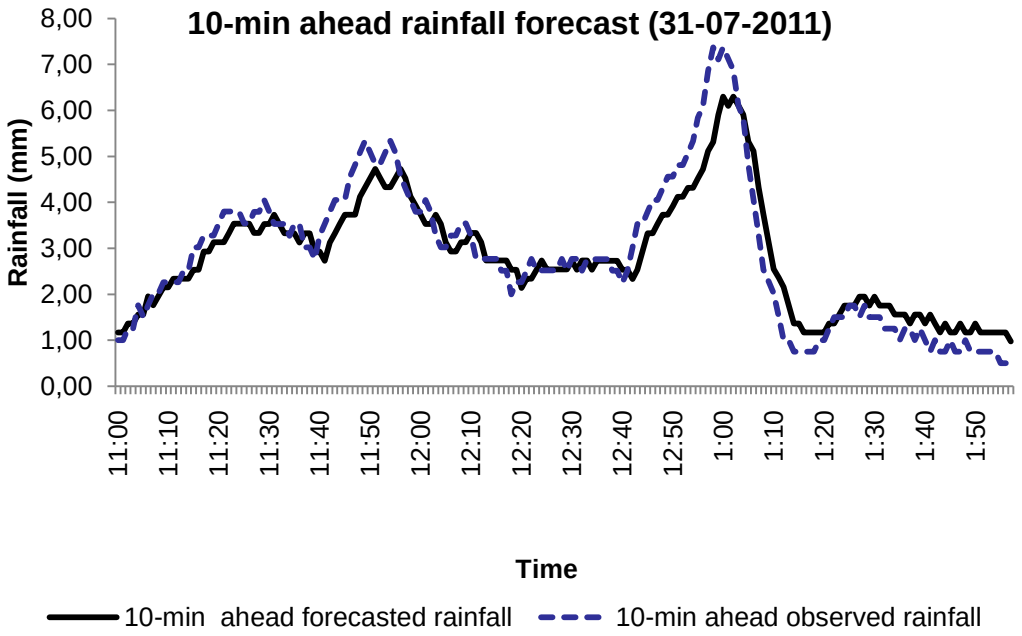
Time series and scatter plot of 5-min ahead rainfall forecast

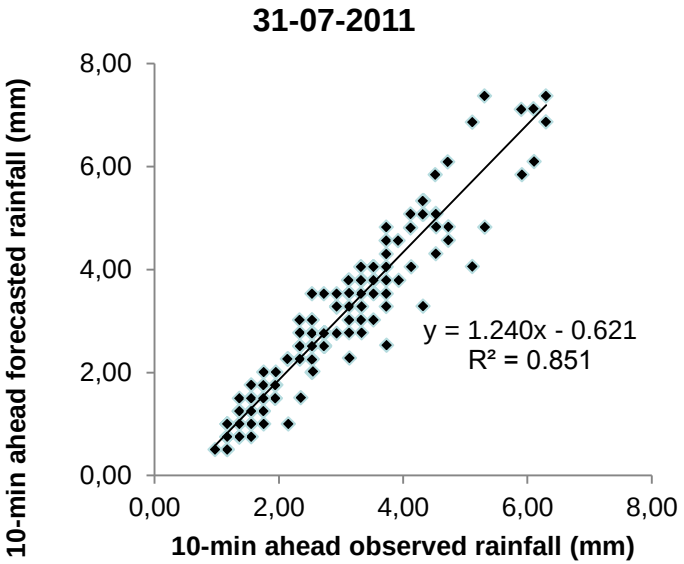




IIT Bombay

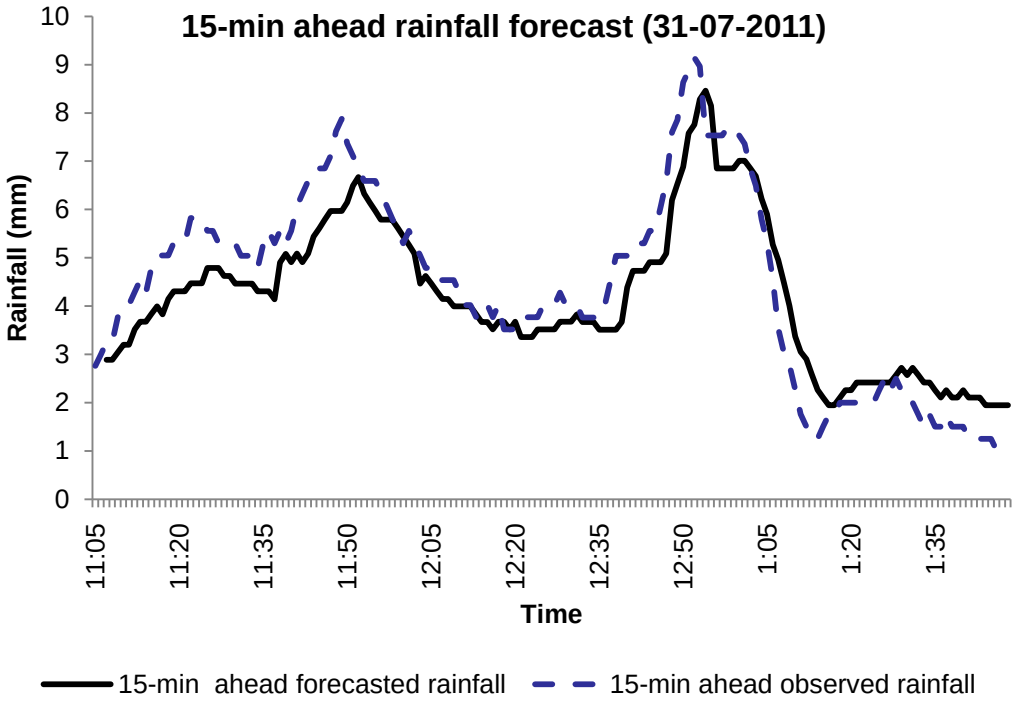
Time series and scatter plot of 10-min ahead rainfall forecast

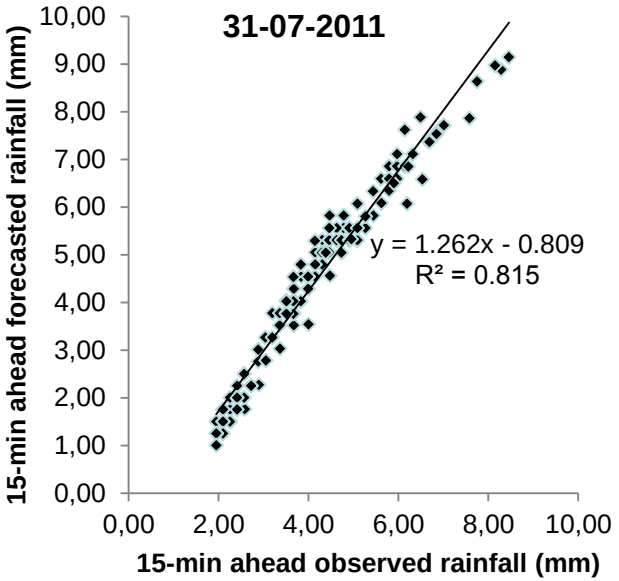




IIT Bombay

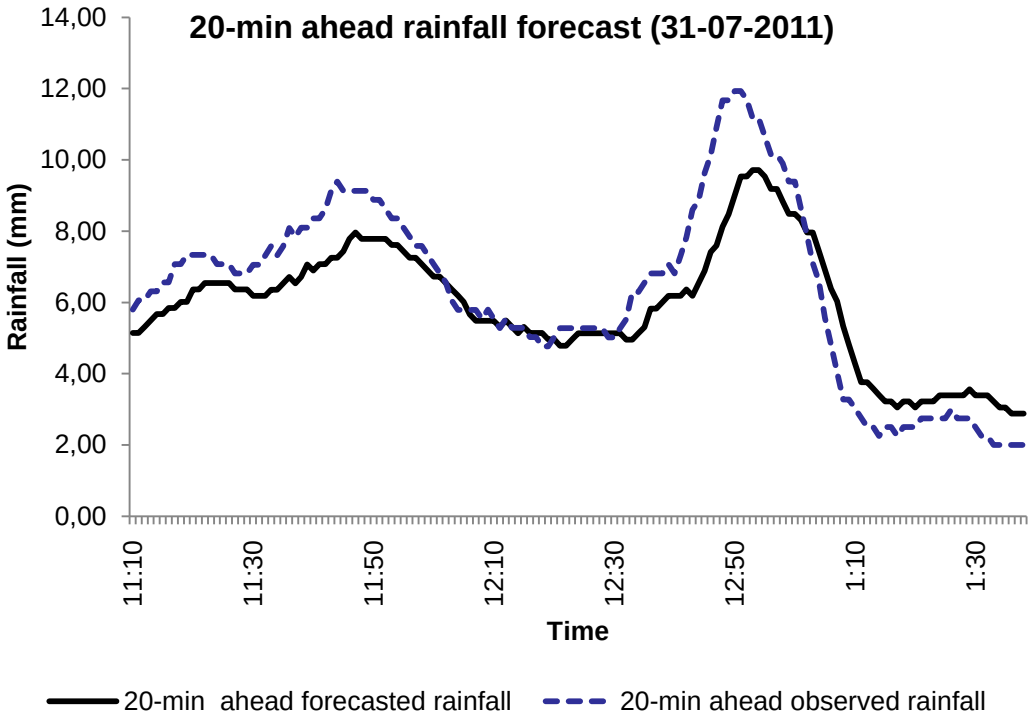
Time series and scatter plot of 15-min ahead rainfall forecast

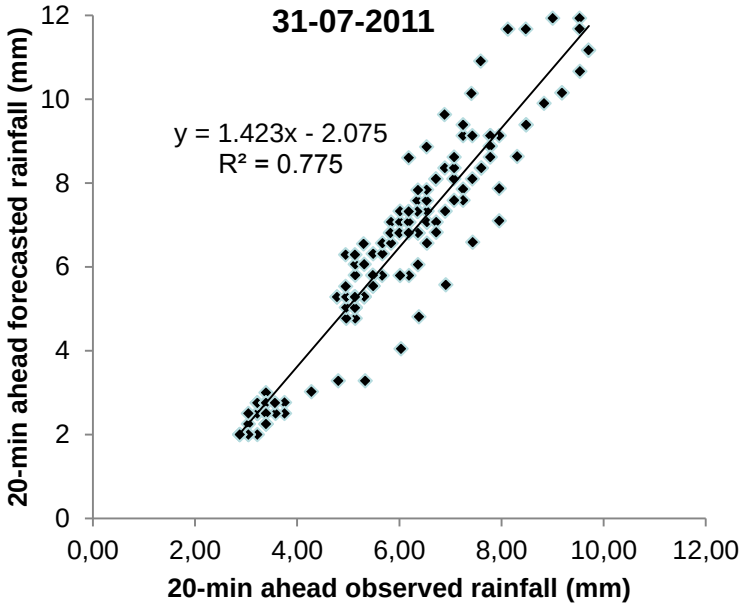




IIT Bombay

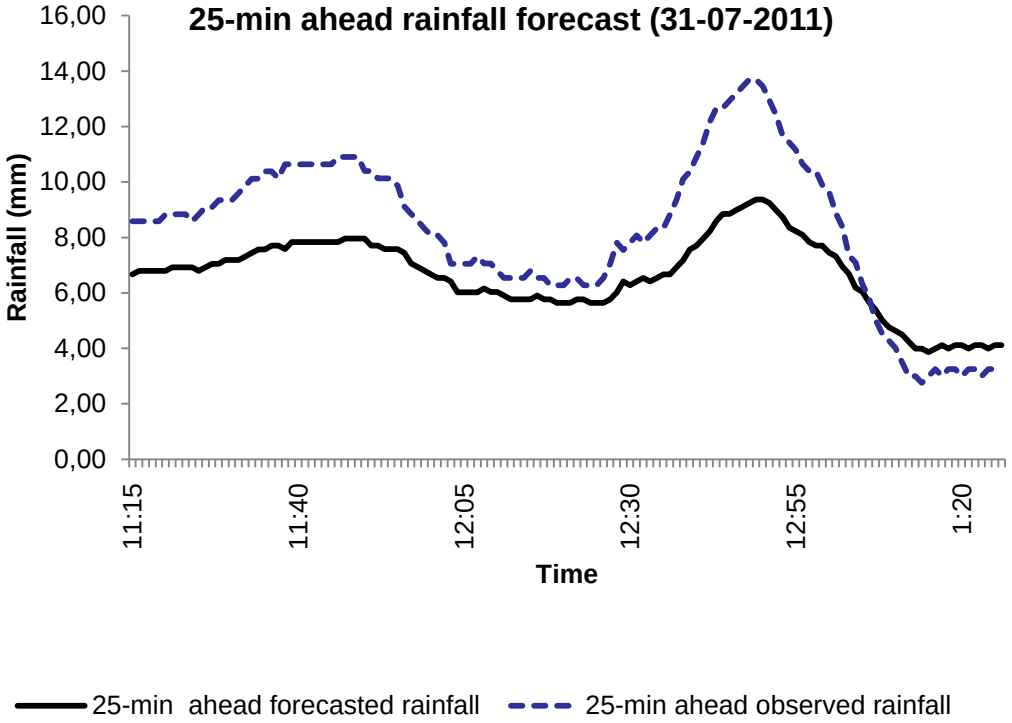
Time series and scatter plot of 20-min ahead rainfall forecast

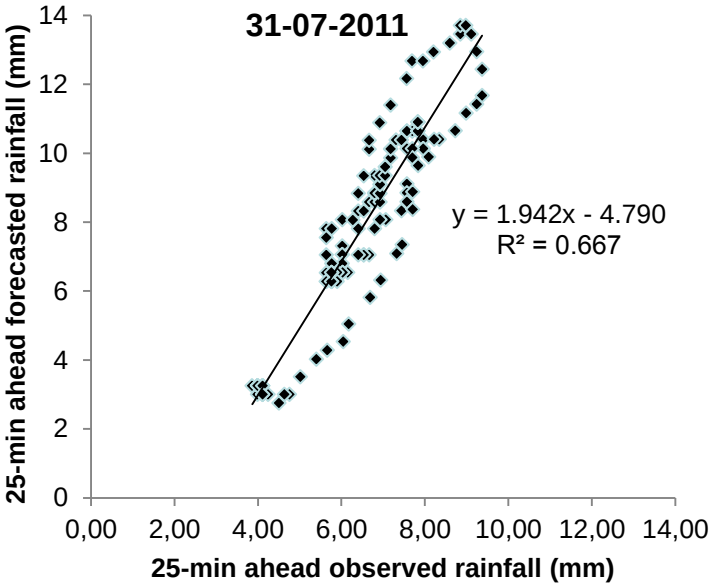




IIT Bombay

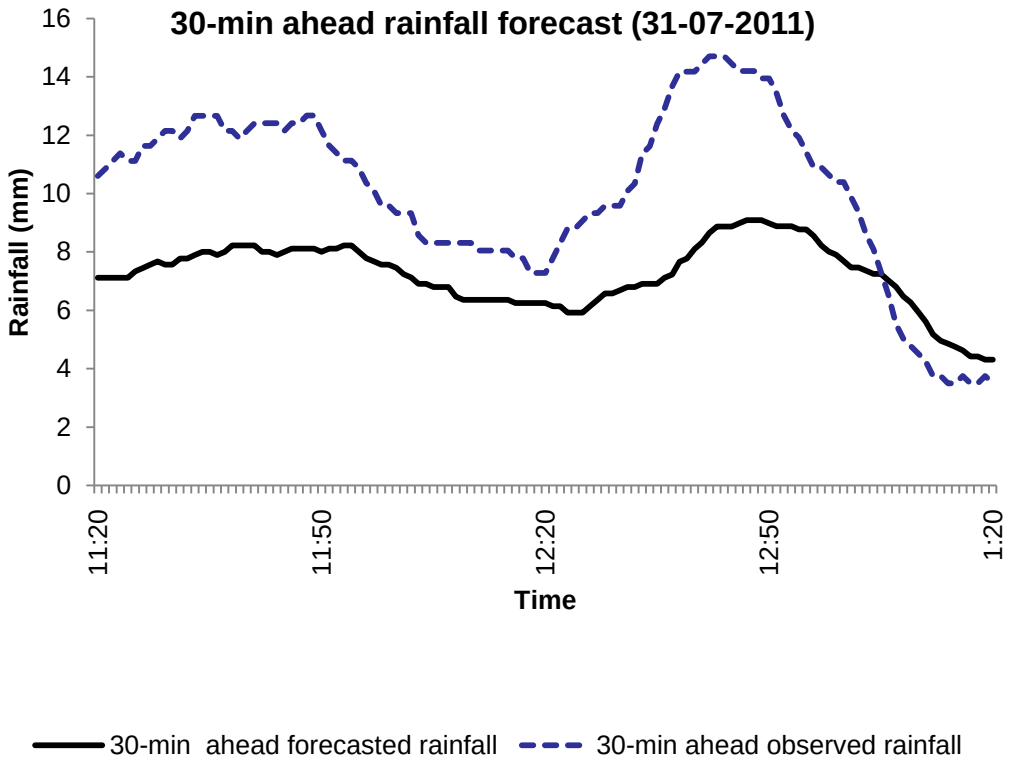
Time series and scatter plot of 25-min ahead rainfall forecast

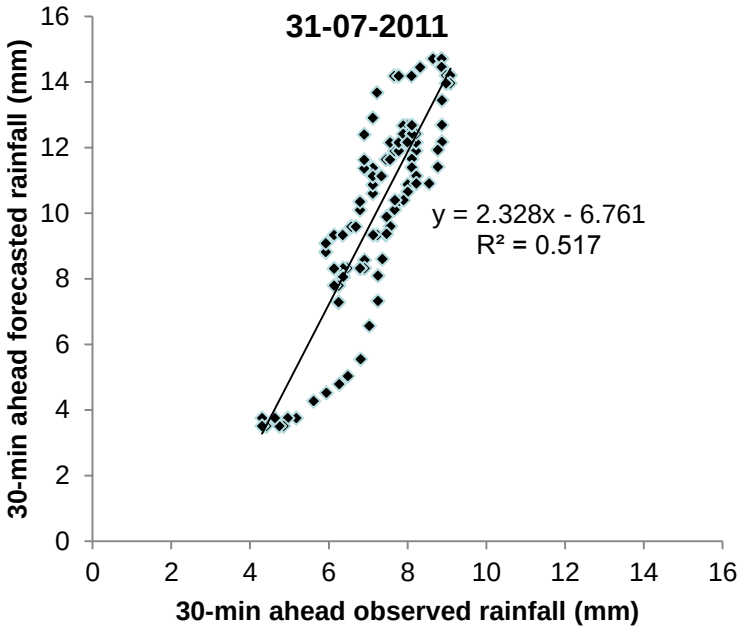




IIT Bombay

Time series and scatter plot of 30-min ahead rainfall forecast





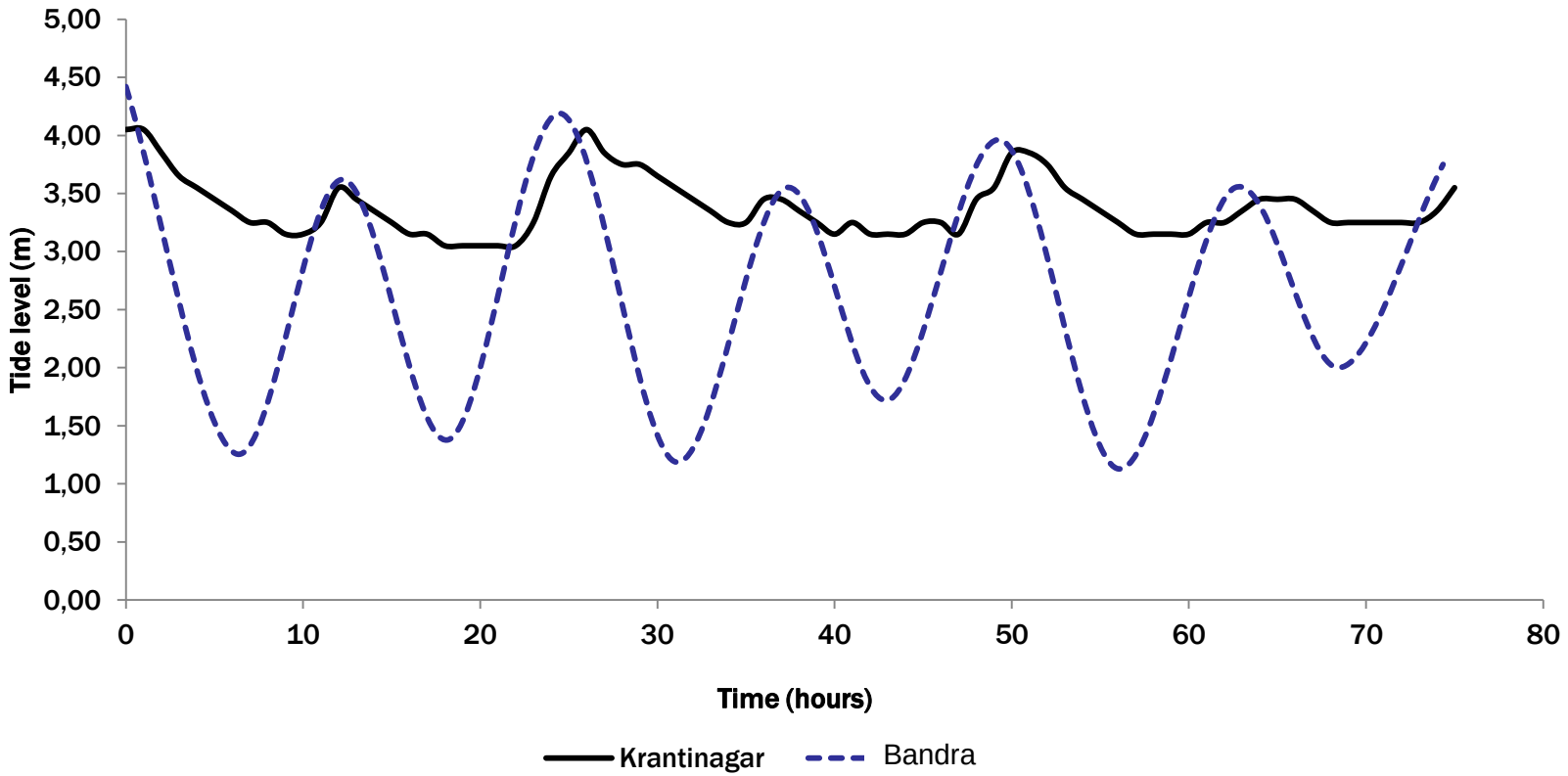
IIT Bombay

Results and discussions

1. Developed rainfall forecast using support vector machine tool can forecast rainfall upto a lead-time of 15-min after which forecast accuracy decrease drastically
2. Forecast model has not able to successfully capture the rainfall peak
3. Rainfall forecast model has shown improvement with segregation of rainfall inputs for training

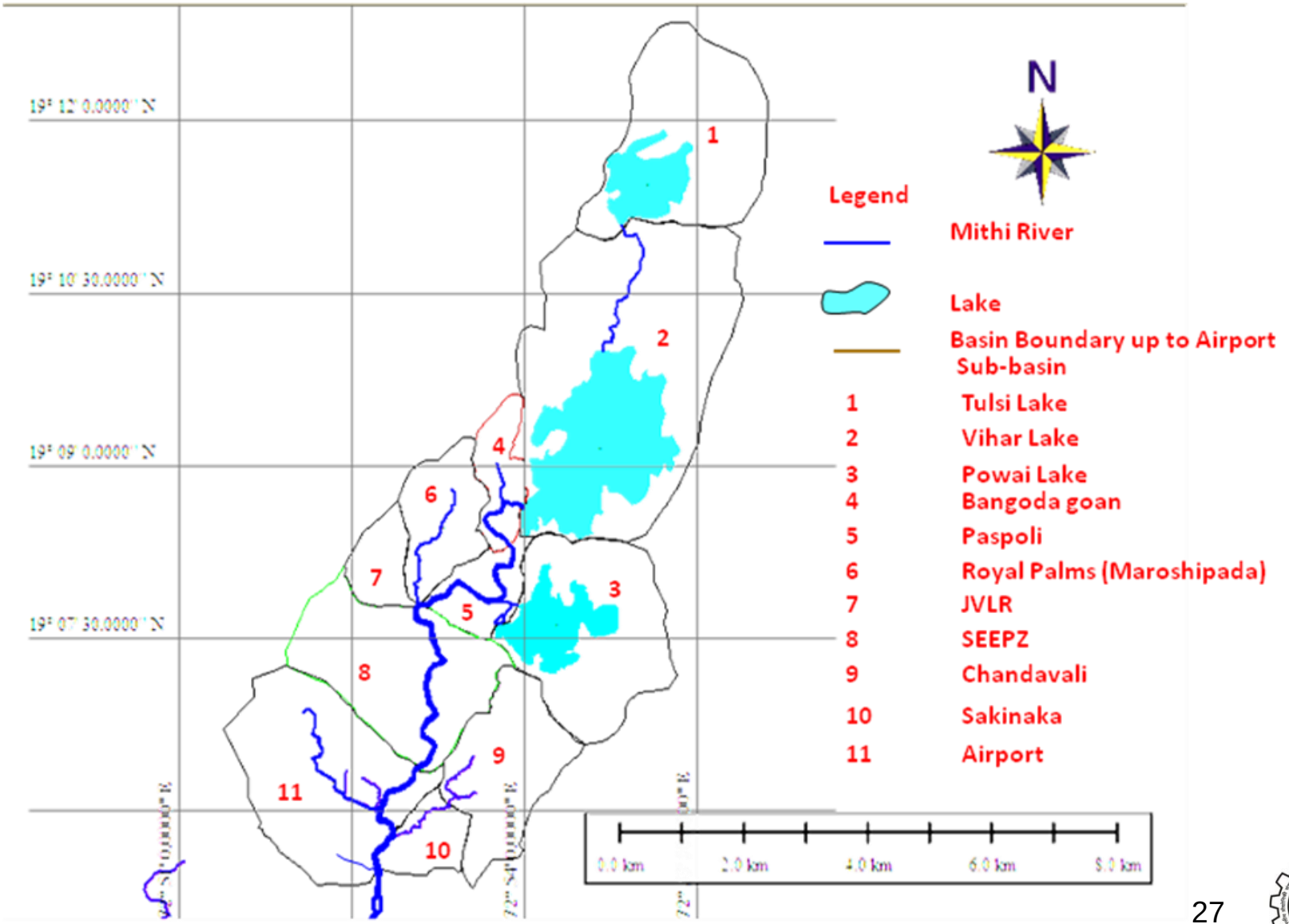
Thank you

Tide levels at Bandra and Krantinagar gauging site (16:00 pm 17-06-10 to 20:00 pm 20-06-10)



IIT Bombay

Catchment contributing to Krantinagar Culvert



IIT Bombay

Accuracy of SVM

Accuracy of SVM influenced by:

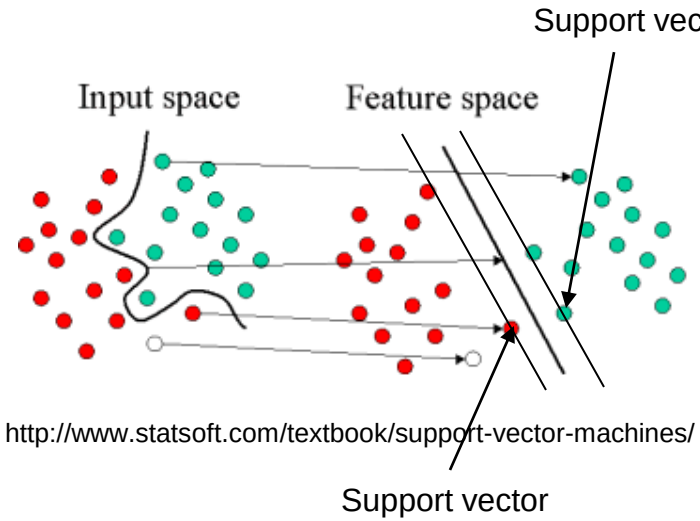
1. Selection of optimal kernel function type

2. Selection of Kernel function parameter viz σ for radial basis function

3. Value of a soft margin constant γ penalty parameter for SVM training and testing stages

Kernels

Kernel is a function that transforms the input data to a high dimensional feature space



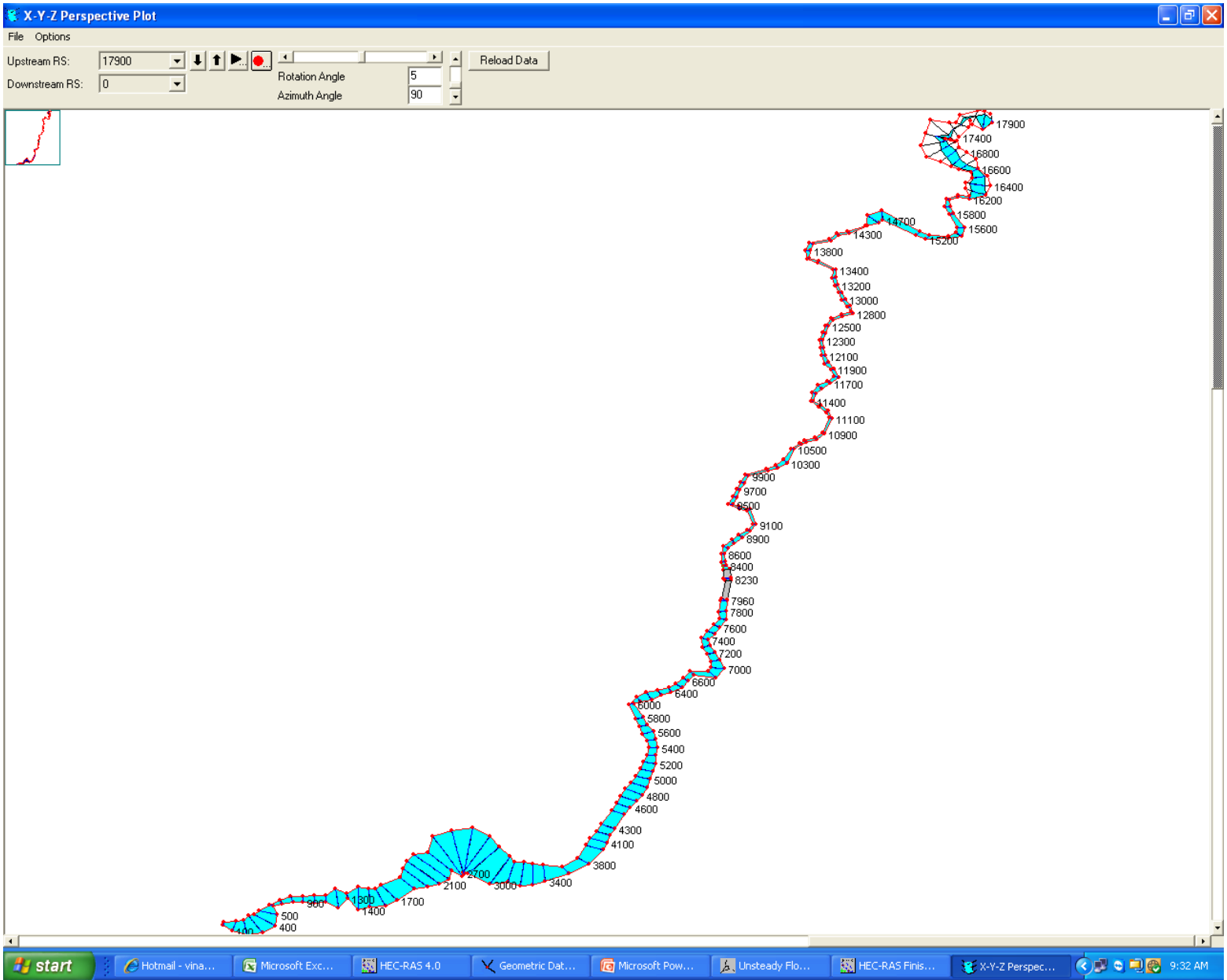
http://www.statsoft.com/textbook/support-vector-machines/

Kernel functions	Equations
Linear	$K(x,y) = x \cdot y$
Polynomial	$K(x,y) = (x \cdot y + 1)p$
Radial basis function	$K(x,y) = exp(-\|\| x - y \|\|^2/\sigma^2)$
Sigmoid	$K(x,y) = tanh(kx \cdot y - d)$

- Radial basis function has been used for present study because in comparison other kernel functions it is able to shorten the computational training process and improve the generalization performance of SVM (Dibike et al. 2001)

29

IIT Bombay

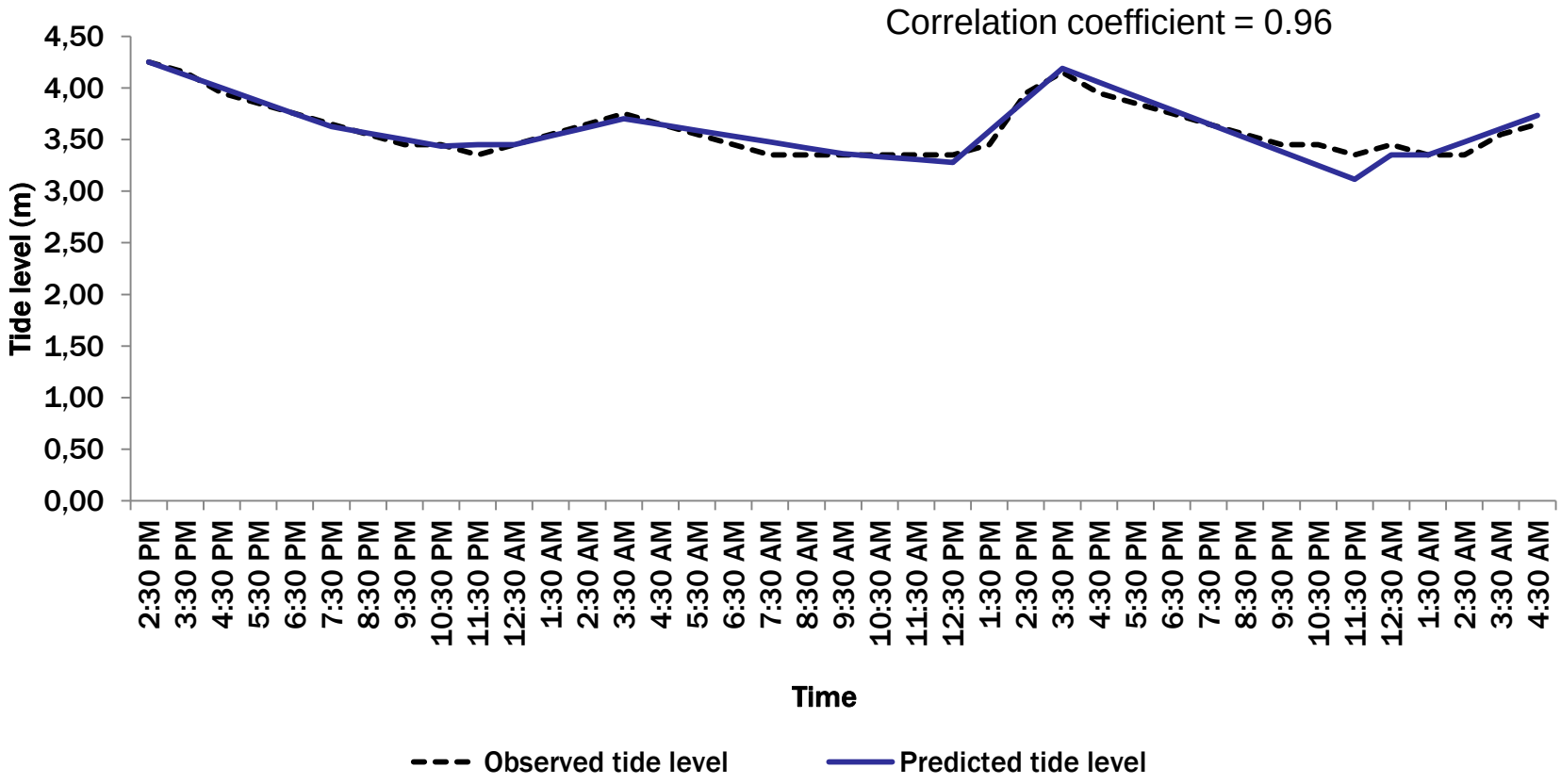

X-Y-Z Perspective Plot
File
Options
Upstream RS:
17900
Downstream RS:
0
Rotation Angle
5
Azimuth Angle
90
Reload Data
17900
17400
16800
16600
16400
16200
15800
15600
15200
14700
14300
13800
13400
13200
13000
12800
12500
12300
12100
11900
11700
11400
11100
10900
10500
10300
9900
9700
9500
9100
8900
8600
8400
8230
7960
7800
7600
7400
7200
7000
6600
6400
6000
5800
5600
5400
5200
5000
4800
4600
4300
4100
3800
3400
3000
2700
2100
1700
1400
1300
900
500
400
start
Hotmail - vina...
Microsoft Exc...
HEC-RAS 4.0
Geometric Dat...
Microsoft Pow...
Unsteady Flo...
HEC-RAS Finis...
X-Y-Z Perspec...
9:32 AM

IIT Bombay

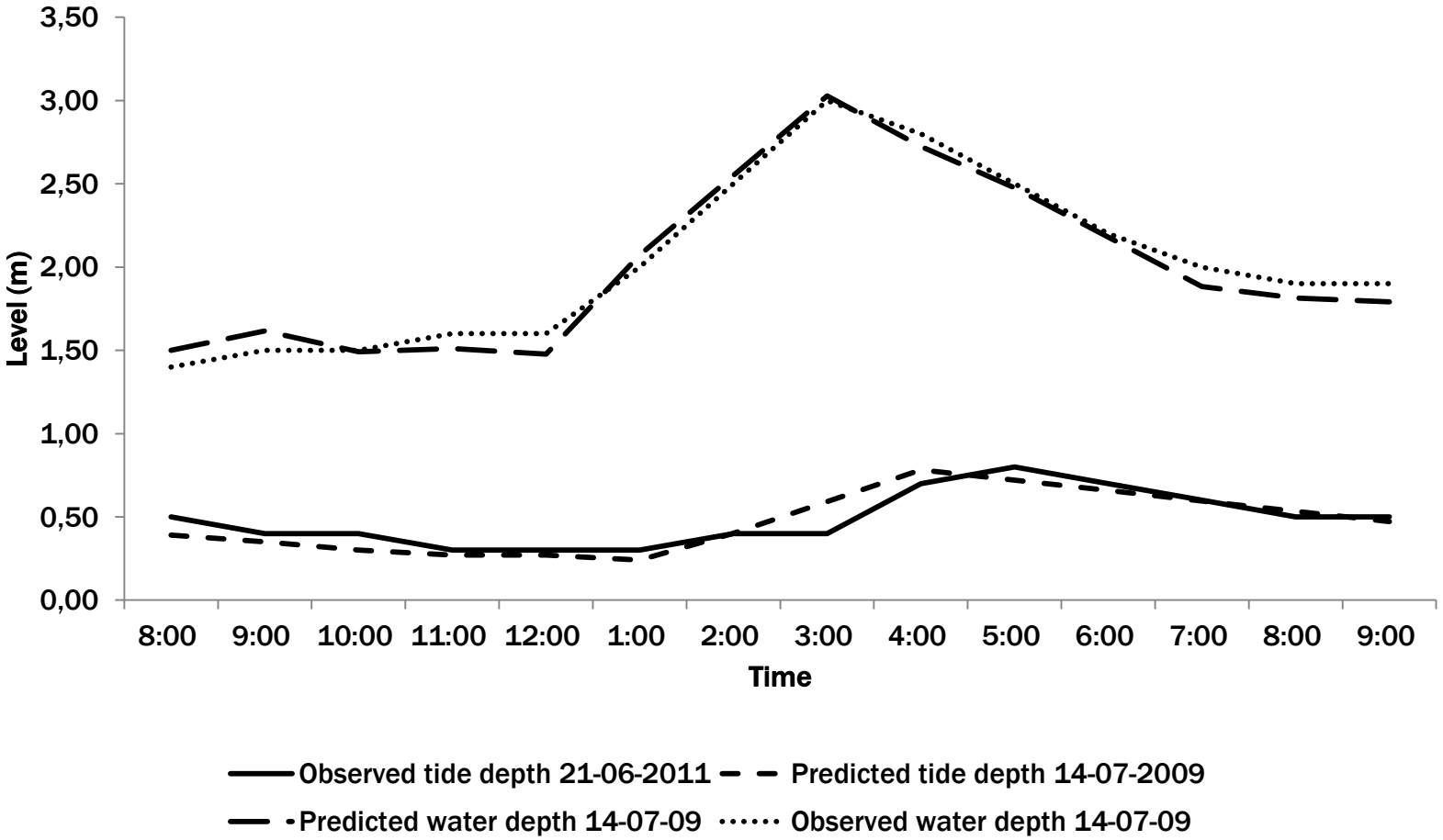
Peak flows from sub-catchments for rainfall intensity of 100 mm/h with runoff coefficient = 0.95

S. No.	Catchment		Area (sq. km)	Flow (m^3/s) for various rainfall intensity (mm/h)						Remark
				(Individual catchment)			Cumulative flow			
	No.	Name		100	120	190	Q_{V100}	Q_{V120}	Q_{V190}	
1	2	Vihar	13.65	200	200	200				Overflow from weir
2	3	Powai Lake	6.221	225	225	225				Overflow from weir
3	4	Bangoda gaon	1.381	36	44	69				
4	5	Passpoli	1.485	39	47	74	500	516	568	Vihar lake + Bangoda gaon + Paspoli + Powai lake (1+ 2 + 3+4)
5	6	Royal Palm	2.376	63	75	119	563	591	687	
6	7	JVLR	1.068	28	34	54	591	625	741	(1+ 2 +3 + 4 +5)
7	8	SEEPZ	5.785	153	183	290	744	808	1031	(1+ 2 +3 + 4 + 5 +6 +7)
8	9	Chandivili	3.868	102	122	194				
9	10	Saki Naka	1.234	33	39	62	135	161	256	Saki Naka + Chandivili (8 + 9)
10	11	Kranti Nagar and Airport	6.701	177	213	336	1056	1182	1623	Total
Total area contributing to Krantinagar culvert = 43.769 sq. km										

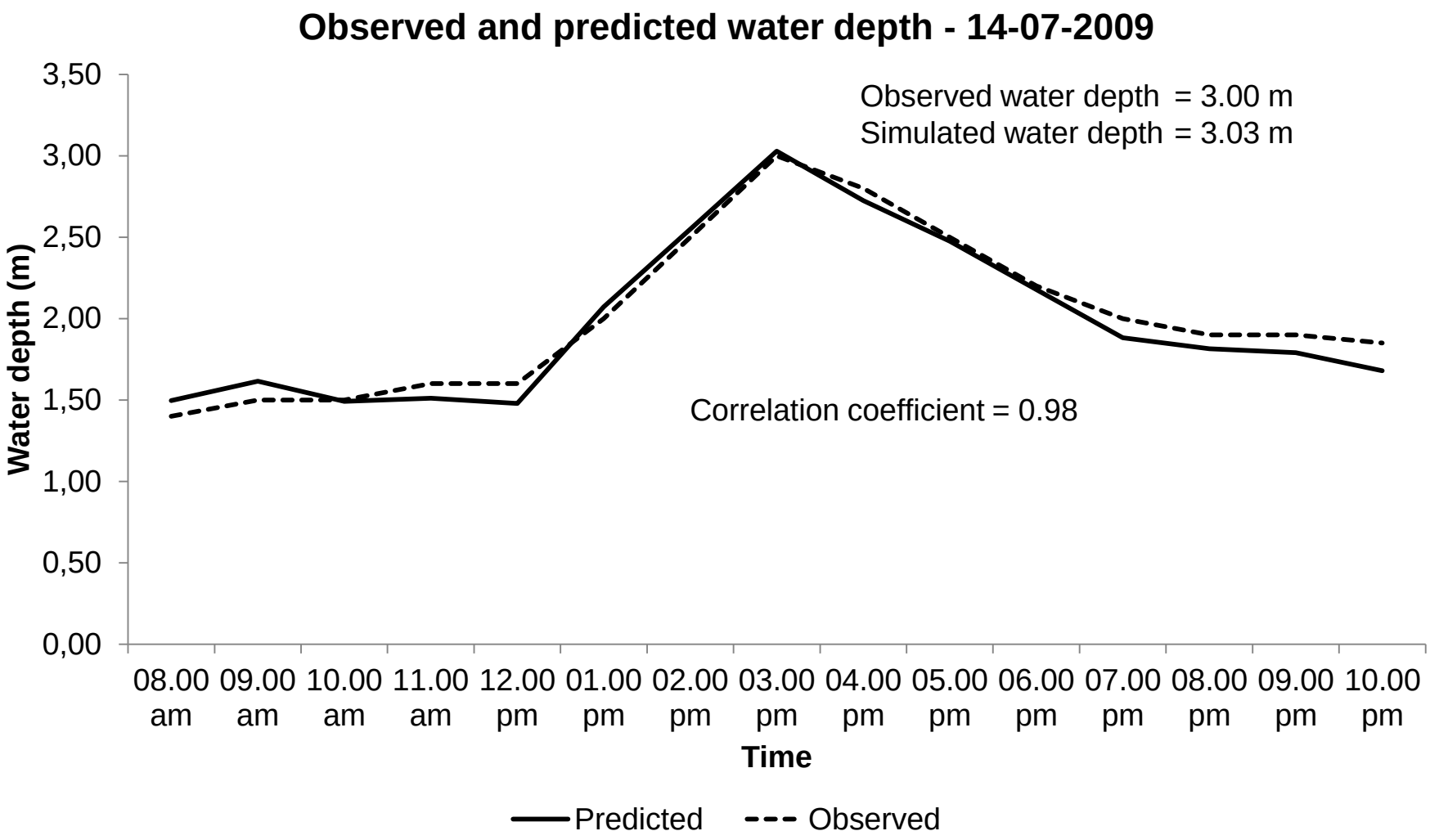
Tidal model calibration - Observed and predicted tide levels at Krantinagar gauge site (2:30 PM 17-06-11 to 4:30 AM 19-06-11)



IIT Bombay

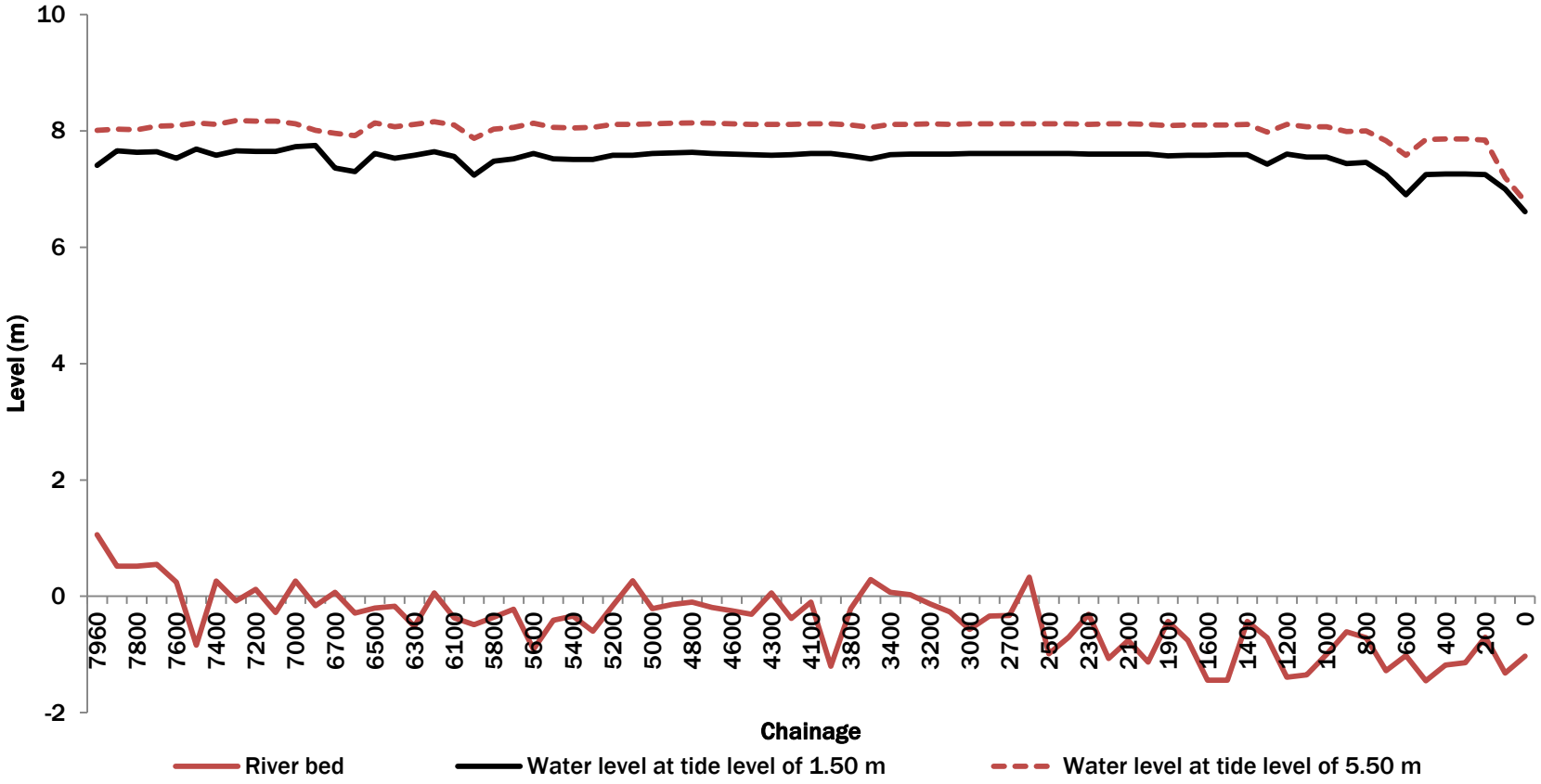

3,50
3,00
2,50
2,00
1,50
1,00
0,50
0,00
Level (m)
8:00
9:00
10:00
11:00
12:00
1:00
2:00
3:00
4:00
5:00
6:00
7:00
8:00
9:00
Time
Observed tide depth 21-06-2011
Predicted tide depth 14-07-2009
Predicted water depth 14-07-09
Observed water depth 14-07-09

IIT Bombay


Observed and predicted water depth - 14-07-2009
Observed water depth = 3.00 m
Simulated water depth = 3.03 m
Correlation coefficient = 0.98
Water depth (m)
3,50
3,00
2,50
2,00
1,50
1,00
0,50
0,00
08.00 am
09.00 am
10.00 am
11.00 am
12.00 pm
01.00 pm
02.00 pm
03.00 pm
04.00 pm
05.00 pm
06.00 pm
07.00 pm
08.00 pm
09.00 pm
10.00 pm
Time
Predicted
Observed

IIT Bombay

Water levels - 1000 m^3/s flow



IIT Bombay

Performance indicators used to measure rainfall forecast accuracy

1. Correlation coefficient (CC) =

$$\frac{\sum_{i=1}^{n}(a_i - \bar{a}) * (f_i - \bar{f})}{\sqrt{\sum_{i=1}^{n}(a_i - \bar{a})^2} * \sqrt{\sum_{i=1}^{n}(f_i - \bar{f})^2}}$$

2. Root mean square error (RMSE) =

$$\sqrt{\frac{\sum_{i=1}^{n}(a_i - f_i)^2}{n}}$$

where,

a_i and f_i represent the actual and forecast rainfall, and $\bar{a}$ and $\bar{f}$ represent the actual and forecast mean rainfall

Rainfall Data

Year	S. No.	Date	Duration (h:min)	Rainfall (mm)	Maximum 15 min rainfall intensity (mm/h)	Remark
2007	1	20/06/2007	1h 50	26.50	37.50	Training
	2	23/06/2007	2h 00	93.50	127.00	
	3	24/06/2007	17:40	115.00	48.75	
	4	28/06/2007	2:10	25.50	34.50	
	5	30/06/2007	14:00	389.75	113.75	
	6	03/07/2007	5:50	53.00	44.75	
	7	27/07/2007	10:00	95.50	105.50	
	8	29/07/2007	2:50	62.00	65.00	
	9	30/07/2007	8:55	40.50	30.50	
	10	07/08/2007	6:50	108.50	54.00	
	11	20/09/2007	4:30	88.00	46.75	

Rainfall Data (Cont..)

Year	S. No.	Date	Duration (h:min)	Rainfall (mm)	Maximum 15 min rainfall intensity (mm/h)	Remark
2008	1	06/06/2008	3:30	40.50	33.50	Training
	2	07/06/2008	4:50	34.00	40.75	
	3	08/06/2008	2:30	39.00	31.50	
	4	11/06/2008	6:45	57.25	39.50	
	5	30/06/2008	3:00	45.00	42.75	
	6	01/07/2008	7:45	125.00	72.25	
	7	10/07/2008	3:10	46.25	52.75	
	8	25/07/2008	2:10	25.25	39.50	
	9	26/07/2008	8:20	64.75	30.50	
	10	27/07/2008	2:30	28.00	49.75	
	11	27/07/2008	21:45	161.5	41.50	
	12	29/07/2008	2:45	24.75	33.50	
	13	6/08/2008	4:30	46.40	37.50	
	14	08/08/2008	1:15	34.50	55.00	
	15	11/08/2008	11:10	63.00	27.50	

Rainfall Data (Cont..)

Year	S. No.	Date	Duration (h:min)	Rainfall (mm)	Maximum 15 min rainfall intensity (mm/h)	Remark
2009	1	26/06/2009	3:10	42.00	45.75	Training
	2	04/07/2009	9:40	117.75	44.75	
	3	13/7/2009	7:40	59.50	31.50	
	4	14/07/2009	7:00	109.50	68.00	
	5	14/07/2009	9:00	75.00	28.25	
	6	21/08/2009	2:10	37.75	66.00	
	7	03/09/2009	5:50	88.75	82.25	
	8	06/10/2009	4:00	40.50	26.45	

Rainfall Data (Cont..)

Year	S. No.	Date	Duration (h:min)	Rainfall (mm)	Maximum 15 min rainfall intensity (mm/h)	Remark
2010	1	14/06/2010	18:20	86.00	83.50	Testing
	2	16/06/2010	11:30	52.25	80.50	
	3	18/06/2010	2:15	40.30	91.25	
	4	22/06/2010	5:15	55.75	55.75	
	5	24/06/2010	3:15	27.50	27.45	
	6	24/06/2010	3:15	28.75	27.25	
	7	25/06/2010	1:35	30.75	47.50	
	8	02/07/2010	4:50	52.00	63.00	
	9	03/07/2010	9:00	76.25	77.00	
	10	03/07/2010	2:50	35.75	52.75	
	11	07/07/2010	3:00	26.50	59.45	
	12	08/07/2010	1:45	29.50	49.75	
	13	08/07/2010	3:50	25.25	25.00	
	14	13/07/2010	3:25	57.50	39.50	
	15	20/07/2010	3:00	30.00	26.50	
	16	24/07/2010	9:40	53.50	35.50	
	17	31/07/2010	4:15	49.50	56.75	
	18	07/08/2010	5:00	41.30	33.50	
	19	14/08/2010	2:45	26.25	26.50	
	20	18/08/2010	2:45	27.30	30.50	
	21	27/08/2010	1:40	29.00	45.00	
	22	28/08/2010	1:30	40.00	78.00	
	23	29/08/2010	6:30	46.50	33.50	

High Intensity Rainfall

In the present study, rainfall > 50 mm/h is defined as a high intensity rainfall which is occurring in very short time (< 1 hour results from :

1. Severe thunderstorms
2. Depressions
3. Cyclones

Support Vector Machines verses Artificial Neural Networks (Vapnik, 1995)

SVM	ANNs
i. Uses structural risk minimization	i. Uses empirical risk minimization
ii. Solution to an SVM is global and unique	ii. Suffer from multiple local minima
iii. SVM training finds a global minimum, and their simple geometric interpretation provides fertile ground for further investigation	iii. Follow a heuristic path, with applications and extensive experimentation preceding theory
iv. SVMs automatically select their model size (by selecting the support vectors)	iv. Computational complexity depend on the dimensionality of the input space
v. Number of hidden nodes are determined by the number of support vectors extracted by the algorithm	v. Use trial and approach for determining the number of hidden nodes

Support Vector Machines

1. Uses structural risk minimization (Vapnik, 1995)
2. Solution to an SVM is global and unique (Vapnik,1998; Burges, 1998)
3. SVMs automatically select their model size (by selecting the Support vectors) (Rychetsky, 2001)
4. Number of hidden nodes is determined by the number of support vectors extracted by the algorithm (Khan and Coulibaly, 2006)

ERM is a measured mean error rate on the training set
Structural risk minimisation:
- Data set is divided into different subset
- Further subset is selected with optimal complexity
- SRM is a selection of classifier that minimize the sum of empirical risk and VC dimension.
- VC dimension is the size of the largest data set that can be shattered by a set of function
- VC dimension is a representation of the capacity of classifier

Data scaling

$$X_n = \frac{X_0 - X_{min}}{X_{max} - X_{min}}$$

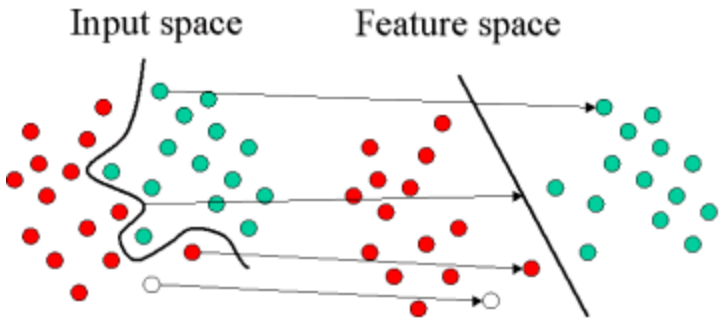
Where X_n = the new value after scale

X_0 = the actual value before scale

X_{min} = the minimum value

X_{max} = the maximum value

The parameter *C* controls the trade off between errors of the SVM on training data and margin maximization (*C* = ∞ leads to hard margin SVM).
Rychetsky (2001), page 82
"The parameter *C* controls the trade-off between the margin and the size of the slack variables."

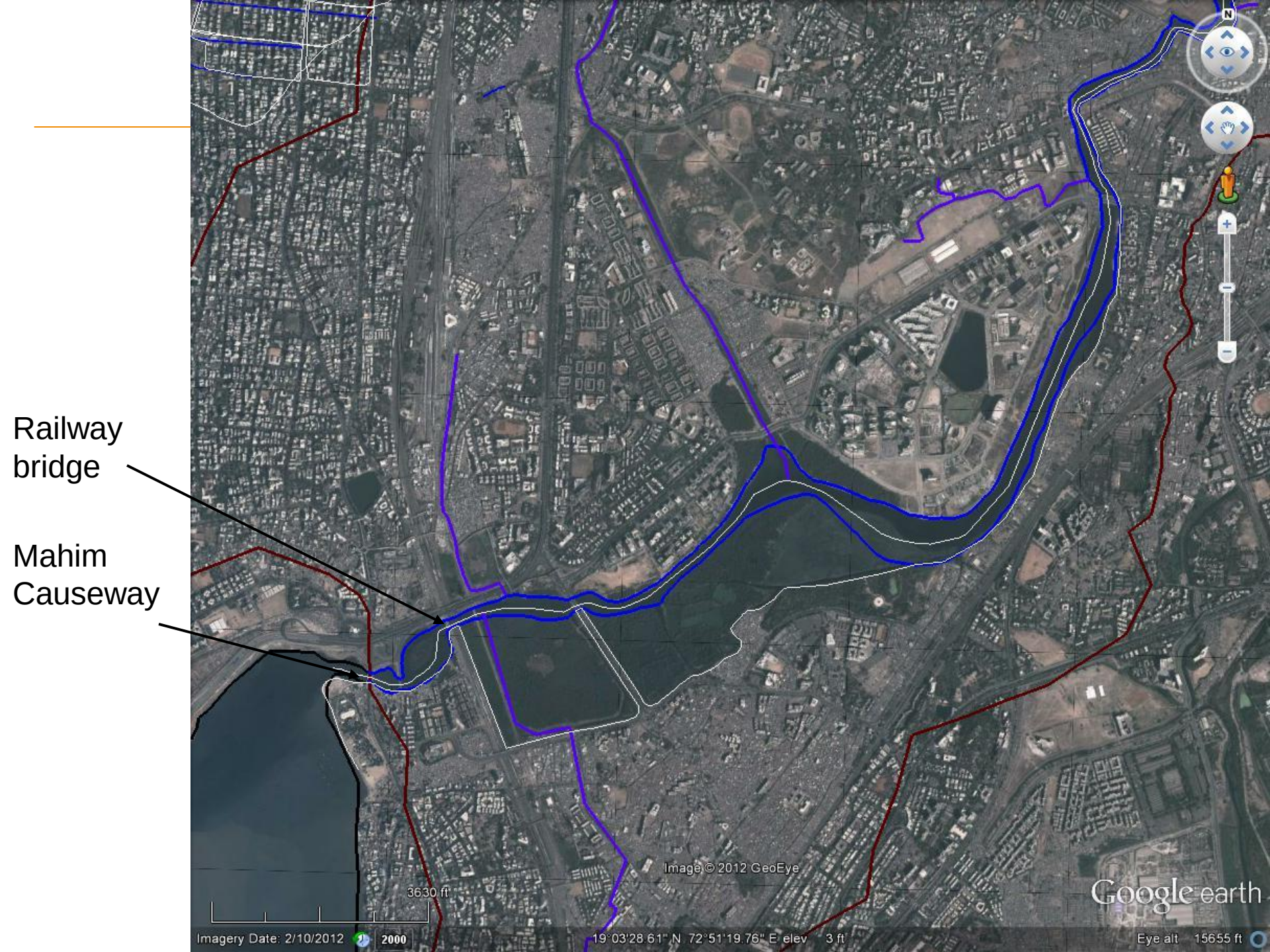


http://www.statsoft.com/textbook/support-vector-machines/

Performance statistics for different lead times

S. No.	Lead time	CC	R^2	RMSE
1	5-min	0.9170	0.8409	0.53
2	10-min	0.8912	0.7942	1.16
3	15-min	0.8626	0.7441	1.25
4	20-min	0.7757	0.6017	2.52
5	25-min	0.7609	0.5790	2.83
6	30-min	0.6760	0.4570	4.42

1. R^2 of 15-min, 10-min and 5-min ahead forecast model are 0.7441, 0.7942 and 0.8409 respectively
2. It is seen that the higher lead time, i.e., (30-min, 25-min and 20-min) gave low accuracy of forecast with R^2 values 0.4570, 0.5790 and 0.6017 respectively. This may be because of decrease in data sets for longer storm durations and high intensities (>50 mm/h = 25 nos.; >70 mm/h = 14 nos.; >100 mm/h = 7 nos.)
3. RMSE value increases from 0.53 to 4.42 from 5-min to 30-min forecast, respectively


Railway bridge
Mahim Causeway
Image © 2012 GeoEye
3630 ft
Imagery Date: 2/10/2012
2000
19°03'28.61" N 72°51'19.76" E elev 3 ft
Eye alt 15655 ft
Google earth