The implementation details of all models are listed as follows:

1) FNN: A two-layer FNN with a hidden size of 128 and using the ReLU function between the two layers as activation.

2) LSTM/GRU: The LSTM and GRU are implemented in a sequence-to-sequence manner to recursively generate multi-step-ahead output. The encoder and decoder follow the same structure, each contains 2 layers of LSTM (GRU) with 64 hidden units. An FNN is applied to the output of the RNN at each time step to convert it to the final prediction. Models are trained using the teacher-forcing strategy with a ratio of 0.5. The learning rate is set to 0.01 with a decaying ratio of 0.1 in 5, 20, and 40 epochs.

3) STGCN: Two spatiotemporal convolutional (ST-Conv) blocks are stacked, followed by an output layer containing two temporal CNN and one FNN. The channels of three layers in ST-Conv blocks are (64, 32, 1) and (64, 32, 128), respectively. Both the graph convolution kernel size and temporal convolution kernel size are set to 3. Similar to our study, the Chebyshev polynomials approximation is used for graph convolution. The learning rate is set to 0.001 with a decaying ratio of 0.7 in every 5 epochs.

4) DCRNN: DCRNN is implemented in a S2S manner to recursively generate multi-step-ahead output. The encoder and decoder follow the same structure, each contains 2 layers of diffusion convolutional GRUs with 64 hidden units. The dual random walk approach is adopted for the diffusion process. The learning rate is set to 0.01 with a decaying ratio of 0.1 in 5, 20, and 40 epochs.

5) ASTGCN: ASTGCN builds three attention-based STGCN for three types of temporal heads (i.e., close, period, and trend) respectively, and fuses their outputs using a parametric-matrix-based weighting function. The hidden units for graph convolution and temporal convolution are set to 64 with a kernel size of 3. The learning rate is set to 0.0001.

6) GWNET: GWNET stacks 2 spatial-temporal layers. Each spatial-temporal layer is constructed by a graph convolution layer and a gated temporal convolution layer. The hidden units for all convolution networks are set to 32 with a kernel size of 2. 1×1 convolution with output channels of 256 is set for skip connection. Two 2D CNNs are stacked as the final output layers by first projecting the channels to 512 and then downsampling to the output dimension.

7) AGCRN: The AGCRN consists of an encoder and a 2D CNN which is used to replace the decoder. The encoder is constructed by 2 layers of adaptive zone-specific graph convolution GRU with 64 hidden units. The dimension of zone embedding is set as 10, and the order of Chebyshev polynomials is set as 2. The learning rate is set to 0.003 with a decaying ratio of 0.75 in 5, 15, 30, and 40 epochs.

8) GMAN: First, the node2vec algorithm is used to learn a zone embedding vector with a dimension of 64 and combined it with temporal embedding vectors. For node2vec, the number of random walks is set as 100 with 50 iterations. Then, an encoder and a decoder both with 2 ST-Attention blocks, one transform attention layer (number of attention heads is set as 2), and 2 FNNs are constructed to generate the final output. The learning rate is set to 0.001 with a decaying ratio of 0.7 when the loss does not decrease for 5 epochs.

9) MTGNN: In MTGNN, temporal convolution and graph convolution are interleaved with each other to capture temporal and spatial dependency respectively. The channel of convolutional layers is set as 32, and the number of layers is set as 3. The temporal inception layer consists of four filter sizes, viz. 1×2, 1×3, 1×6, and 1×7. The output module consists of two 1×1 convolution layers. MTGNN also includes a graph self-learning module with a zone embedding dimension of 40.

10) STGODE: The hidden dimensions of temporal dilation convolution blocks are set to 64, 32, 64, and 3 Spatial-Temporal Graph ODE blocks are contained in each layer. The regularized hyperparameter is set to 0.8. The thresholds of the spatial adjacency matrix are set to 10 and 0.5 respectively, and the threshold of the semantic adjacency matrix is set to 0.6. The model is trained using Adam optimizer with a learning rate of 0.01.

11) STG-NCDE: The order of Chebyshev polynomials is set to 2 and the zone embedding size is set to 8. The dimensionality of the hidden vector is set to 64. The learning rate is set to 0.001 and the weight decay is 0.001.

12) Multi-ATGCN: The Adam optimizer is employed to minimize the model loss with the learning rate decaying (Starting from 0.003, decaying by 75% once the number of epochs reaches 5, 10, 20, and 30 epochs). Each model is run 50 epochs and an early stop strategy with a patience of 10 is used by monitoring the loss in the validation set. The batch size is set as 16. Gradient clipping (maximum norm = 5) is performed during the training process to mitigate exploding gradients. Dropout (ratio = 0.1) is applied before the output layer.