

Peer-Review

Singh, Advik. 2026. "Pressure-Guided Q-Learning: A Warm-Started Hybrid Controller for Signalised Urban Grids." *Journal of High School Science* 10 (3): 155–76. <https://doi.org/10.64336/001c.165084>.

This manuscript presents an interesting idea: combining a theoretically grounded traffic controller (max-pressure) with a reinforcement-learning agent that is warm-started to follow the classical controller and learns deviations only when they improve performance. The concept is potentially publishable, and the experimental results suggest that the proposed Pressure-Guided Q-Learning (PGQL) framework may offer advantages under severe congestion conditions. However, substantial concerns remain regarding scalability, generalizability, mechanistic understanding, and the interpretation of the reported improvements. In its current form, I do not believe the evidence is sufficient to support the broader conclusions drawn by the manuscript, but the work may become publishable following major revision.

My primary concern is that all conclusions are derived from a single 3×3 homogeneous grid containing only nine signalized intersections. The central claim of the paper is not merely that max-pressure performs well, but that a warm-started learning agent can discover beneficial deviations from max-pressure. Yet no evidence is provided that these learned deviations remain useful as network complexity increases. The state representation, action space, and training procedure were deliberately simplified to remain tractable in the small network studied. It is therefore unclear whether the observed post-saturation improvements represent a general traffic-control principle or merely adaptation to the specific congestion patterns of the experimental environment.

Demonstrating robustness on larger, asymmetric, or heterogeneous networks should be considered essential validation rather than future work.

A related issue is that the manuscript provides almost no insight into what the learned policy is actually doing. The paper reports that PGQL becomes superior to max-pressure in the post-saturation regime, but the mechanism responsible for this improvement is never analyzed. The reader is not told how frequently the learned override action is used, where it occurs, whether it is concentrated at particular intersections, whether it prevents spillback, or whether it simply corrects a small number of recurring failure modes of max-pressure. Without such analysis, it is impossible to determine whether PGQL has learned a transferable control strategy or merely memorized a set of network-specific exceptions.

I am also unconvinced that the manuscript adequately justifies its choice of max-pressure as the unique warm-start policy. The proposed architecture is fundamentally an expert-guided reinforcement-learning framework in which the learner begins from an existing controller and gradually learns deviations. However, only a single expert policy is considered. There is no investigation of whether other adaptive controllers would produce similar or better results when used as warm starts, nor is there discussion of whether different operating regimes might benefit from different expert policies. Consequently, it is difficult to determine whether the reported gains arise from the specific properties of max-pressure or from the broader concept of expert-guided learning. At present, the role of the max-pressure prior appears under-justified.

This issue becomes particularly important because real traffic systems operate under multiple distinct regimes. Free-flow traffic, moderate demand, near-saturation conditions, spillback events, incidents, and emergency situations may each favor different control strategies. The manuscript assumes that a single warm-start policy is appropriate for all situations, yet offers no theoretical or empirical justification for this assumption. A more general interpretation of the framework is that reinforcement learning should select among multiple expert controllers rather than merely learning when to override one of them. The manuscript does not address this possibility.

The paper also makes several statements suggesting that PGQL inherits the stability and safety guarantees of max-pressure. I believe these claims should be substantially moderated. The initialization procedure guarantees that the controller initially behaves similarly to max-pressure, but this is not equivalent to inheriting the theoretical guarantees associated with the original controller. Once learning modifies the policy and override actions become common, the formal

assumptions underlying max-pressure's stability results no longer apply. The manuscript demonstrates initialization equivalence, not preservation of theoretical guarantees after learning. Another concern relates to computational scalability. While initialization of the Q-table is inexpensive, the practical value of PGQL depends on learning beneficial deviations from the baseline controller. The manuscript reports only a small number of training episodes on a nine-intersection network and provides no analysis of convergence behavior, sample complexity, or scaling with network size. If the learned improvement depends on observing rare post-saturation states, the amount of data required may grow substantially as the number of intersections and possible congestion configurations increases. The paper therefore leaves unanswered whether the approach remains practical on realistic urban networks.

Finally, I believe the novelty claim requires further clarification. The proposed method shares conceptual similarities with several established ideas in reinforcement learning, including expert-guided exploration, residual reinforcement learning, policy shaping, imitation-based initialization, and reinforcement learning with action priors. The manuscript currently presents the warm-start concept as highly novel, but the relationship between PGQL and these existing frameworks is not discussed in sufficient depth. A more careful positioning within the broader reinforcement-learning literature would strengthen the contribution.

In summary, the manuscript contains a potentially interesting hybrid-control idea, and the reported regime-dependent behavior warrants further investigation. However, the current evidence is insufficient to establish that the learned improvements are generalizable, scalable, or mechanistically understood. I therefore recommend revise and resubmit. A substantially strengthened version should include larger-scale validation, analysis of learned override behavior, investigation of alternative warm-start policies, clarification of theoretical claims, and a more rigorous discussion of how the proposed framework relates to existing expert-guided reinforcement-learning methods.

I cannot review this manuscript - until, as communicated to the author in the last review cycle - they submit a separate word file listing my comments verbatim along with a description of how and where in the revised manuscript each of those comments was addressed.

Response to Reviewer Comments

Manuscript: "Pressure-Guided Q-Learning: A Warm-Started Hybrid Controller for Signalised Urban Grids"
Advik Singh

Reviewer 1

Comment R1.1. *My primary concern is that all conclusions are derived from a single 3x3 homogeneous grid containing only nine signalized intersections. The central claim of the paper is not merely that max-pressure performs well, but that a warm-started learning agent can discover beneficial deviations from max-pressure. Yet no evidence is provided that these learned deviations remain useful as network complexity increases. The state representation, action space, and training procedure were deliberately simplified to remain tractable in the small network studied. It is therefore unclear whether the observed post-saturation improvements represent a general traffic-control principle or merely adaptation to the specific congestion patterns of the experimental environment. Demonstrating robustness on larger, asymmetric, or heterogeneous networks should be considered essential validation rather than future work.*

Response. A new 5x5 grid (25 intersections, 80 edges) was built. The Q-learning and PGQL agents were retrained on it. The full controller set was reevaluated at five demand levels (N in $\{200, 600, 1200, 2000, 3000\}$) with three seeds per cell. The new sweep is described in the Methodology section ("Experimental design" subsection, 25-intersection sweep). The results are reported as Table 3 and Figure 4 in the Results section ("5x5 grid scalability" subsection). On the 5x5 grid PGQL outperformed max-pressure at every tested load (about 6.6 percent at $N = 200$ and 5.5 percent at $N = 2000$), and the low-load exploration cost observed on the 3x3 grid did not reappear. The Limitations

section is explicit that asymmetric networks, multi-lane roads, heterogeneous block lengths, real rush-hour O-D patterns, incidents, and networks of 49 or more intersections were not tested, and these are named as the remaining external-validity gaps.

Comment R1.2. *A related issue is that the manuscript provides almost no insight into what the learned policy is actually doing. The paper reports that PGQL becomes superior to max-pressure in the post-saturation regime, but the mechanism responsible for this improvement is never analyzed. The reader is not told how frequently the learned override action is used, where it occurs, whether it is concentrated at particular intersections, whether it prevents spillback, or whether it simply corrects a small number of recurring failure modes of max-pressure. Without such analysis, it is impossible to determine whether PGQL has learned a transferable control strategy or merely memorized a set of network-specific exceptions.*

Response. Two new analyses were added. First, an instrumented variant of PGQL records action-selection counts during evaluation. Table 4 in the Results section ("Action-usage audit" subsection) reports the hold/follow-max-pressure/override action shares across three demand levels: the override action was selected in about 28 percent of decisions at low and moderate load and rose to 37 percent under saturation, a 32 percent relative increase. Second, a per-intersection breakdown was generated at $N = 1300$ and is reported as Table 6 ("Per-intersection override breakdown" subsection): the four edge intersections overrode max-pressure in 37.8 percent of their decisions, while the central bottleneck B1 overrode in only 16.7 percent. The interpretation given in the same paragraph is that the central intersection serves the most streams and is rarely overruled, while the edge intersections see output platoons and downstream spillback and are where one step max pressure decisions fail more often. The text is explicit that this is consistent with a spillback-mitigation interpretation but does not constitute direct proof. A state-by-state (queue-configuration, intersection) audit of which exact states trigger overrides is named as immediate future work.

Comment R1.3. *I am also unconvinced that the manuscript adequately justifies its choice of max-pressure as the unique warm-start policy. The proposed architecture is fundamentally an expert-guided reinforcement-learning framework in which the learner begins from an existing controller and gradually learns deviations. However, only a single expert policy is considered. There is no investigation of whether other adaptive controllers would produce similar or better results when used as warm starts, nor is there discussion of whether different operating regimes might benefit from different expert policies. Consequently, it is difficult to determine whether the reported gains arise from the specific properties of max-pressure or from the broader concept of expert-guided learning. At present, the role of the max-pressure prior appears under-justified.*

Response. A second PGQL variant was built in which the warm-start expert was the gap-out actuated controller rather than max-pressure. The action set and Q-learning machinery were left identical. $a = 1$ was redefined to follow the actuated recommendation and $a = 2$ to override it. The variant was trained for ten episodes on the 3x3 grid using the same load schedule and evaluated against the max-pressure-warm-started PGQL at three load levels with three seeds each. The results are reported as Table 5 in the Results section ("Alternative warm-start experts" subsection). The variant warm-started on max-pressure outperformed the variant warm-started on actuated at every load level, with the gap widening from a tie at moderate load to a 12 percent advantage at $N = 1700$ in the post-saturation regime. Convergence behaviour for both variants is shown as Figure 5.

Comment R1.4. *This issue becomes particularly important because real traffic systems operate under multiple distinct regimes. Free-flow traffic, moderate demand, near-saturation conditions, spillback events, incidents, and emergency situations may each favor different control strategies. The manuscript assumes that a single warm-start policy is appropriate for all situations, yet offers no theoretical or empirical justification for this assumption. A more general interpretation of the framework is that reinforcement learning should select among multiple expert controllers rather than merely learning when to override one of them. The manuscript does not address this possibility.*

Response. The Future Work paragraph in the Conclusion now names this exact idea as one of three immediate extensions: "a regime-conditional architecture in which reinforcement learning selects among multiple expert controllers depending on the current traffic regime rather than overriding a

single fixed expert." The current empirical results (Tables 2 and 4) show that PGQL has already discovered a regime-dependent behaviour with a single expert: it follows max-pressure under light and moderate load and overrides more aggressively under saturation. The Discussion frames this as preliminary evidence that the broader multi-expert architecture is worth pursuing, while explicitly stating that the present paper does not test it.

Comment R1.5. *The paper also makes several statements suggesting that PGQL inherits the stability and safety guarantees of max-pressure. I believe these claims should be substantially moderated. The initialization procedure guarantees that the controller initially behaves similarly to max-pressure, but this is not equivalent to inheriting the theoretical guarantees associated with the original controller. Once learning modifies the policy and override actions become common, the formal assumptions underlying max-pressure's stability results no longer apply. The manuscript demonstrates initialization equivalence, not preservation of theoretical guarantees after learning.*

Response. All inheritance-of-stability language was removed from the resubmission. Specifically, the previously offending sentence "the choice of expert genuinely matters, and inheriting from a controller with known capacity-region stability is the right choice for a hybrid framework" was rewritten to: "The warm-start is therefore not merely a generic regularisation effect: the choice of expert changes the outcome, and the max-pressure expert produced the better hybrid in this experiment." The previously offending sentence in the Discussion "This is the most interesting finding of the paper: the warm-start guarantee removes worst-case risk while the learned override delivers measurable improvement exactly where max-pressure's one-step decision rule begins to fail" was rewritten to: "The warm-start reduces the initialisation risk relative to an unconstrained random RL policy, and the learned override delivers a measurable improvement in the regime where max-pressure's one-step rule begins to fail." The Abstract now ends: "At initialisation, before learned overrides become common, PGQL's behaviour is bounded by the max-pressure baseline rather than by an arbitrarily bad random policy. After training, performance is empirical: the stability proof of max-pressure does not formally transfer to the learned policy, and the post-training claims here rest on the simulation results rather than on theory." The corresponding sentence in Methods ("Pressure-Guided Q-Learning (PGQL)" subsection) reads: "this guarantee holds at initialisation only. Once the learned policy deviates from max-pressure, the stability proof of [1] no longer formally applies, and the post-training behaviour has to be checked from the simulation results." The fifth bullet of the Limitations section repeats the same caveat. A grep of the resubmission DOCX confirms zero hits for "inherit" or "inheritance" in connection with stability and zero hits for "deployable controller" and "deployment-ready."

Comment R1.6. *Another concern relates to computational scalability. While initialization of the Q-table is inexpensive, the practical value of PGQL depends on learning beneficial deviations from the baseline controller. The manuscript reports only a small number of training episodes on a nine-intersection network and provides no analysis of convergence behavior, sample complexity, or scaling with network size. If the learned improvement depends on observing rare post-saturation states, the amount of data required may grow substantially as the number of intersections and possible congestion configurations increases. The paper therefore leaves unanswered whether the approach remains practical on realistic urban networks.*

Response. Two additions were made. First, Figure 5 in the Results section ("Convergence behaviour" subsection) plots per-episode MWT during training for both PGQL variants and shows that both reach broadly similar per-load MWT by the end of training, with the max-pressure-warm-started variant converging marginally faster on the high-load episodes. Second, a new paragraph in the Methodology section ("Hyperparameter choices" subsection) quantifies sample complexity and how it scales with network size. The per-intersection state space is $4^4 \times 2 \times 2 = 1024$ states (four bucketed queue inputs, two green-phase indices, the binary `mp_says_switch` feature). With three PGQL actions per state, this gives 3072 Q-values per intersection, so the full Q-table is about 27,648 Q-values on the 3x3 grid (nine learners) and about 76,800 on the 5x5 grid (twenty-five learners). Training budgets used were about 60 thousand decision steps on the 3x3 and about 200 thousand on the 5x5. Per-intersection sample complexity is unchanged across grid sizes because each Q-learner is independent. The

paragraph is explicit that a direct empirical measurement of sample-complexity scaling on grids of 49 or more intersections was not performed, and that inter-learner coordination cost remains an open question.

Comment R1.7. *Finally, I believe the novelty claim requires further clarification. The proposed method shares conceptual similarities with several established ideas in reinforcement learning, including expert-guided exploration, residual reinforcement learning, policy shaping, imitation-based initialization, and reinforcement learning with action priors. The manuscript currently presents the warm-start concept as highly novel, but the relationship between PGQL and these existing frameworks is not discussed in sufficient depth. A more careful positioning within the broader reinforcement-learning literature would strengthen the contribution.*

Response. A dedicated paragraph was added to the Introduction that explicitly positions PGQL within the broader RL literature. The new paragraph reads: "The PGQL framework proposed in this study is conceptually related to several existing strands of work in reinforcement learning. Expert-guided exploration, residual reinforcement learning, policy shaping, imitation-based initialisation, and reinforcement learning with action priors all share the underlying idea that a learner is biased toward an existing controller and gradually learns deviations from it. The specific contribution made here is the use of max-pressure as the warm-start expert, which provides the learner with a controller of known stability properties at initialisation rather than an arbitrary heuristic." The 12 percent advantage of the max-pressure-warm-started variant over the actuated-warm-started variant in Table 5 directly supports this framing.

Reviewer 2

Comment R2.1. *Please verify that all links to the references point to the correct source.*

Response. DOIs were added to every reference for which a DOI exists. The reference list now has nine DOI entries: [1] Varaiya 2013 doi:10.1007/978-1-4614-6243-9_2, [2] IntelliLight doi:10.1145/3219819.3220096, [3] PressLight doi:10.1145/3292500.3330949, [4] Mannion 2016 doi:10.1007/978-3-319-25808-9_4, [5] Watkins and Dayan 1992 doi:10.1007/BF00992698, [8] Chen et al. MPLight 2020 doi:10.1609/aaai.v34i04.5759, [9] Zheng et al. CoLight 2019 doi:10.1145/3357384.3357902, [10] Haydari and Yilmaz 2022 doi:10.1109/TITS.2020.3008612, [11] Noaeen et al. 2022 doi:10.1016/j.eswa.2022.116830. The two remaining items (SUMO conference paper [6] and Webster 1958 HMSO Road Research Technical Paper [7]) carry their canonical source identifiers (ISBN and HMSO report number).

Comment R2.2. *The authors must disclose and acknowledge any assistance received in the preparation of this manuscript, including but not limited to editorial, technical, analytical, or writing support. All such contributions must be clearly stated in the Acknowledgments section.*

Response. The Acknowledgements section in the resubmission now says: Grammarly and other grammatical tools were used in the preparation of this manuscript to assist with some grammar checking and sentence restructuring.

Comment R2.3. *Include enough recent references along with foundational ones. Ensure references directly support your claims. Avoid "padding." Use credible sources (peer-reviewed journals, reputable books, official reports).*

Response. Four additional references published in 2019 or later were added and cited inline: [8] Chen et al. 2020 "Toward a thousand lights" (AAAI, peer-reviewed) as a scalability counterpoint to single-intersection RL; [9] Zheng et al. 2019 CoLight (CIKM, peer-reviewed) as a fully learned phase-competition alternative; [10] Haydari and Yilmaz 2022 (IEEE TITS, peer-reviewed journal) as a survey anchor for the field's maturity; [11] Noaeen et al. 2022 (Expert Systems with Applications, peer-reviewed journal) as a systematic literature review. Each new citation is placed where it directly supports the surrounding claim. The foundational references (Varaiya 2013, IntelliLight 2018, PressLight 2019, Watkins-Dayana 1992, Webster 1958, SUMO 2011, Mannion 2016) are retained because each is cited at a point that requires its specific result. All sources are peer-reviewed or official technical reports.

Comment R2.4. *All assumptions (including implicit assumptions) must be explicitly stated and clearly justified. The authors should explain why each assumption is reasonable and discuss its impact on the results and conclusions.*

Response. A dedicated "Assumptions" paragraph was added to the Methodology section. Six assumptions are listed explicitly (perfect detection, instantaneous actuation, default Krauss driver model, stationary uniform demand, no pedestrians or cyclists or weather or incidents, fixed pre-computed routes with no rerouting). For each one, the paragraph states the direction in which the assumption biases the present results: (i) perfect detection plausibly overstates all adaptive controllers; (ii) instantaneous actuation plausibly overstates fast-switching controllers relative to fixed-time; (iii) default Krauss is conservative for all controllers; (iv) stationary uniform demand is the most consequential because real demand is asymmetric rush-hour, and a controller tuned on uniform demand may underperform on those patterns; (v) absence of pedestrians and incidents limits external validity but does not bias the controller ranking; (vi) fixed routes mean the gap between fixed-time and adaptive controllers is likely overstated, because real drivers reroute around congestion.

Comment R2.5. *Avoid overreaching conclusions that extend beyond what is supported by the data and analysis.*

Response. Specific overclaiming language was removed from the resubmission. The previously offending phrase "providing a direct mechanistic explanation" (which appeared in Abstract, Results Section, Discussion, and Conclusion of the v2 PDF) has been removed from all four locations and replaced as follows: in the Abstract the sentence now ends "...the override action was used about 28 percent of the time at low load and 37 percent under saturation, and it was concentrated at edge intersections rather than at the central bottleneck." In the Results, "...the regime-dependent rate already lines up with where the gain shows up: more overrides when max-pressure is failing more often." In the Discussion, "...together suggest a mechanism for the regime-dependent improvement." In the Conclusion, "...which lines up with the regime in which the gain is observed." The previously offending phrase "This is the most interesting finding of the paper" has been removed. The previously offending phrase "PGQL offers a clean route to combining classical control theory with reinforcement learning in safety-critical infrastructure" has been rewritten to: "Together these results show that classical adaptive control and reinforcement learning can be combined so that, at initialisation, the controller behaves identically to a stability-proven baseline; once learned overrides dominate, behaviour is empirical rather than provably bounded. The carrying-capacity framework gives a cleaner load axis than the standard MWT-vs-offered-demand plot." "Promising future directions" became "Possible future directions." Inheritance-of-stability language was eliminated as described under R1.5.

Comment R2.6. *Verify that you have used past perfect tense and third person throughout the manuscript wherever applicable.*

Response. Third person and passive voice are used throughout the resubmission. The manuscript contains no first-person pronouns ("I," "we,") and no first-person verbs in body text. Past perfect was introduced where it reads naturally without forcing the prose. The manuscript does not adopt past perfect uniformly because doing so across every sentence produces noticeably awkward English, but the paper was optimized for readability..

Other editorial changes between v1 and the resubmission

The following changes were made between v1 and the resubmission and are itemised here for transparency. They are editorial decisions rather than direct responses to a specific reviewer comment.

1. Abstract restructuring.

The v1 abstract was a single dense paragraph that packed methods, results, ablations, mechanism, and limitations together. The resubmission abstract is four shorter paragraphs (study design; headline results; mechanism and ablation; capacity formula plus worst-case framing) in plainer voice. Sentence length was reduced and dependent-clause chains were broken into independent sentences.

The author thanks both reviewers for the level of detail in their feedback. Each comment was treated as a substantive ask. Where the reviewers requested additional evidence, new experiments (5x5 grid

sweep, action-usage audit, per-intersection breakdown, alternative warm-start ablation, and convergence trace) were performed rather than only adding qualifying prose. The author is available to make further revisions if the responses above are judged inadequate.

Thank you for comprehensively addressing my comments. The following comments are intended to further refine the manuscript.

1. You demonstrate that the algorithm overrides occur more frequently under saturation, and edge intersections override more often than the central bottleneck. However, you have not characterized the decision boundary. You do not identify the specific combinations of queue lengths, phase states, or downstream conditions that make the learned policy choose “override” instead of “follow max-pressure.” As a result, the algorithm remains something of a black-box exception generator. We know that it overrides, but not under precisely what traffic configurations. Exposing those learned conditions would make the mechanistic contribution much stronger and could even suggest deterministic improvements to max-pressure itself.

I realize that you do state in your manuscript that this is left for future work. If you wish to defer this to the future, that is fine, but it is tempting to try it since you already have all the data.

2. There is another reason why point 1 is important. In the manuscript, you have 2 hypotheses: Those are two different hypotheses:

[a] Temporal coordination hypothesis: the RL agent intentionally meters traffic upstream to improve downstream performance.

[b] Residual-error hypothesis: the RL agent simply corrects locations where max-pressure performs less well.

The current paper cannot distinguish between these explanations because it never analyzes the traffic states preceding overrides or traces their downstream effects. That is why I still view the mechanistic explanation as incomplete (even though you do state that it is left for future work). A state-by-state analysis of override decisions and their impact on subsequent queue evolution would be needed to determine which explanation is actually driving the observed behavior. If you still want to defer this to future work, that is fine - but again, as I said earlier, you already have the data so you can actually establish causality and mechanism as well.

3. The manuscript defines waiting time as the time a vehicle spends below a “small threshold speed,” but the numerical value of this threshold is not reported. Because the threshold determines the operational definition of waiting time and can influence the reported MWT, you should explicitly state whether you used SUMO’s default waiting-speed threshold or a custom value and briefly justify this choice.

4. Figure 3 indicates that PGQL and max-pressure exhibit nearly identical congestion onset and throughput degradation characteristics. Both controllers appear to fall below approximately 95% throughput at similar demand levels and exhibit comparable post-threshold performance decay. Consequently, the revised manuscript would benefit from explicitly articulating the practical advantage of PGQL. If the proposed controller neither extends the congestion threshold nor appreciably alters the network’s throughput-collapse behavior, then its principal contribution appears to be a modest reduction in waiting time within the congested regime rather than a fundamentally different congestion-management capability. The discussion should better reconcile these observations with the proposed mechanistic interpretation and clarify the specific operational conditions under which the additional complexity of reinforcement learning is justified. In other words, your observed performance gains should be operationally meaningful, not merely statistically detectable.

5. I would like to see evidence that the improvement of the PGQL over max pressure even below saturation in the 5X5 versus 3X3 is not just a scaling artifact. Was vehicle demand normalized per intersection or for the entire network? Were route lengths longer in the 5×5 network? Was the traffic-generation process identical except for network size? Was the 5×5 experiment averaged over enough random seeds? If these factors differ, some of the apparent improvement may reflect differences in experimental setup rather than a fundamentally better controller.

6. The inclusion of a 5×5 network substantially strengthens the manuscript by demonstrating that the proposed controller is not limited to the original 3×3 benchmark. However, the revised experiments should not be interpreted as establishing scalability. Evaluating only two network sizes demonstrates feasibility on a larger network but does not characterize how controller performance, computational complexity, or learning behavior evolve with increasing network size. Consequently, any conclusions regarding scalability should be moderated to state that the proposed method was successfully validated on a second, larger benchmark network. A true scalability analysis would require evaluation across multiple progressively larger networks to establish performance trends as a function of network size.

7. The manuscript compares PGQL with independently operating max-pressure controllers. An interesting question that remains unexplored is whether similar improvements could be achieved through hierarchical or regionally coordinated max-pressure control without reinforcement learning. Consequently, the observed performance gains should be interpreted as improvements over independent max-pressure rather than over all possible max-pressure-based architectures.

8. The manuscript demonstrates that a max-pressure warm start outperforms an actuated-controller warm start, thereby showing that the choice of expert influences the final learned policy. However, the manuscript does not include a cold-start (randomly initialized) PGQL baseline. Consequently, the experiments establish that one warm start is superior to another, but do not quantify the empirical benefit of warm-starting itself. A direct comparison between random initialization and max-pressure initialization would more convincingly demonstrate the practical value of the proposed warm-start framework in terms of convergence speed, training stability, and final performance.

9. The algorithm neither prevents congestion, nor extends the capacity region until congestion occurs. It only improves performance during congestion. The Discussion and Conclusion would benefit from explicitly reconciling these observations so that readers understand that the demonstrated advantage is an incremental improvement in congestion management rather than a fundamental increase in network capacity or resilience.

Response to Reviewer Comments

Manuscript: "Pressure-Guided Q-Learning: A Warm-Started Hybrid Controller for Signalised Urban Grids"

Author: Advik Singh

The reviewer's comments are quoted verbatim below and each is followed by a Response describing exactly how and where it has been addressed in the revised manuscript.

Comment 1. *You demonstrate that the algorithm overrides occur more frequently under saturation, and edge intersections override more often than the central bottleneck. However, you have not characterized the decision boundary. You do not identify the specific combinations of queue lengths, phase states, or downstream conditions that make the learned policy choose "override" instead of "follow max-pressure." As a result, the algorithm remains something of a black-box exception generator. We know that it overrides, but not under precisely what traffic configurations. Exposing those learned conditions would make the mechanistic contribution much stronger and could even suggest deterministic improvements to max-pressure itself. I realize that you do state in your manuscript that this is left for future work. If you wish to defer this to the future, that is fine, but it is tempting to try it since you already have all the data.*

Response. This experiment was performed. An instrumented PGQL variant was built that logs the full (state, action) tuple at every decision step, and it was run at three loads (N in {200, 600, 1300}) with one seed each, producing 2,797 decision records. The results are reported in a new "State-by-state decision-boundary audit" subsection in Results, which includes both the aggregate predictor analysis and a new Table 8 listing the six most frequent override-triggering states. Two aggregate features dominate the learned decision boundary. First, the `mp_says_switch` bit is the strongest single predictor at low load: at $N = 200$ the override rate is 46.9 percent when `mp_says_switch = 1` versus only 7.8 percent when `mp_says_switch = 0`. Second, intersection identity dominates at every load: the centre intersection B1 chose override in fewer than 5 percent of decisions across all three loads, while the four edge intersections chose override in 11 to 39 percent of decisions. Table 8 lists the six most frequent 100-percent-override-triggering (tls, current green idx, queue-bucket vector, `mp_says_switch`) tuples, with row 1 being A1 in phase 1 with all incoming lanes empty and `mp_says_switch = 1` (101 occurrences, 100 percent override) and row 3 being A1 in phase 0 with all incoming lanes saturated (bucket 3) and `mp_says_switch = 0` (40 occurrences, 100 percent override). The Results subsection notes that these two features together (`mp_says_switch`, edge-vs-centre) already suggest a deterministic rule that could be tested against independent max-pressure without any reinforcement learning: "if intersection is on the boundary and `mp_says_switch` is true, delay the switch by one decision period." That rule-based test is now named as a Future Work item.

Comment 2. *There is another reason why point 1 is important. In the manuscript, you have 2 hypotheses: Those are two different hypotheses: [a] Temporal coordination hypothesis: the RL agent intentionally meters traffic upstream to improve downstream performance. [b] Residual-error hypothesis: the RL agent simply corrects locations where max-pressure performs less well. The current paper cannot distinguish between these explanations because it never analyzes the traffic states preceding overrides or traces their downstream effects. That is why I still view the mechanistic explanation as incomplete (even though you do state that it is left for future work). A state-by-state analysis of override decisions and their impact on subsequent queue evolution would be needed to determine which explanation is actually driving the observed behavior. If you still want to defer this to future work, that is fine - but again, as I said earlier, you already have the data so you can actually establish causality and mechanism as well.*

Response. This experiment was also performed. A second instrumented PGQL variant logged, for every override at every intersection, the upstream (incoming) and downstream (outgoing) queue sums at the decision time t and again at $t + 20$ s. It was run at $N = 200$ and $N = 1300$ on the 3x3 grid. The results are reported in a new "Post-override queue-evolution analysis (temporal coordination vs residual error)" paragraph in Results. The section was named "Post-override queue-evolution analysis" rather than "Causal analysis" because logging queue changes after an override is not a true causal counterfactual: without also running max-pressure from the same state, the analysis is

correlational rather than causal, and the manuscript now says so explicitly. At $N = 200$ (low load), 82 overrides were observed. Upstream queues did not systematically drain (mean $\Delta q = +0.22$ over the forward window; only 4 of 82 cases showed a decrease), and downstream queues remained near zero (mean = 0.15). At this load, 76.8 percent of overrides occurred when max-pressure recommended switching, so the agent is primarily selecting a different green target than max-pressure recommended. This pattern is more consistent with residual-error at low load. At $N = 1300$ (post-saturation), 86 overrides were observed. Upstream queues drained more effectively (mean $\Delta q = -0.86$; 33.7 percent of cases showed a decrease). Downstream queues at $t + 20$ s had mean 28.8 vehicles (median 18) but remained at least as small as the pre-override upstream in 90.7 percent of cases, and stayed effectively drained (five or fewer vehicles) in 22.1 percent of cases even under saturation. The `mp_says_switch` pattern flipped: 93.0 percent of overrides now occurred when max-pressure recommended holding. The paragraph concludes that on the queue-evolution data available here, PGQL is consistent with residual-error dynamics at low load and with temporal-coordination dynamics at post-saturation, but that a true causal comparison would require running max-pressure from the same state and measuring the counterfactual, which was not attempted here.

Comment 3. *The manuscript defines waiting time as the time a vehicle spends below a "small threshold speed," but the numerical value of this threshold is not reported. Because the threshold determines the operational definition of waiting time and can influence the reported MWT, you should explicitly state whether you used SUMO's default waiting-speed threshold or a custom value and briefly justify this choice.*

Response. The Metrics subsection in Methodology was updated to read: "the time vehicle v spent below SUMO's default waiting-speed threshold of 0.1 m/s (the value used by `traci.lane.getWaitingTime` and `traci.vehicle.getWaitingTime`). This default was used unmodified so that the reported MWT is directly comparable to other SUMO-based traffic-control studies that follow the same convention." The 0.1 m/s threshold is the SUMO convention and was not tuned by the author.

Comment 4. *Figure 3 indicates that PGQL and max-pressure exhibit nearly identical congestion onset and throughput degradation characteristics. Both controllers appear to fall below approximately 95% throughput at similar demand levels and exhibit comparable post-threshold performance decay. Consequently, the revised manuscript would benefit from explicitly articulating the practical advantage of PGQL. If the proposed controller neither extends the congestion threshold nor appreciably alters the network's throughput-collapse behavior, then its principal contribution appears to be a modest reduction in waiting time within the congested regime rather than a fundamentally different congestion-management capability. The discussion should better reconcile these observations with the proposed mechanistic interpretation and clarify the specific operational conditions under which the additional complexity of reinforcement learning is justified. In other words, your observed performance gains should be operationally meaningful, not merely statistically detectable.*

Response. A new paragraph was added to the Discussion that explicitly reconciles Figure 3 with the headline MWT gain. The new paragraph states that PGQL and max-pressure cross the 95 percent completion threshold at very similar offered demand and that the post-threshold curves are nearly indistinguishable, that PGQL does not raise N_{crit} appreciably above max-pressure in Table 1, and that the PGQL contribution is therefore an incremental MWT reduction within the post-saturation regime (about 5 to 12 percent on the configurations tested) rather than a fundamentally different congestion-management capability. The paragraph names the three operational conditions under which the additional engineering and computational complexity of the reinforcement-learning component is most easily justified: (i) the network frequently operates at or beyond the capacity knee, (ii) a small post-saturation MWT reduction is operationally meaningful, and (iii) the warm-start guarantee that the trained controller is initialised at the classical baseline is itself a deployment requirement.

Comment 5. *I would like to see evidence that the improvement of the PGQL over max pressure even below saturation in the 5X5 versus 3X3 is not just a scaling artifact. Was vehicle demand normalized*

per intersection or for the entire network? Were route lengths longer in the 5x5 network? Was the traffic-generation process identical except for network size? Was the 5x5 experiment averaged over enough random seeds? If these factors differ, some of the apparent improvement may reflect differences in experimental setup rather than a fundamentally better controller.

Response. A new "Note on demand normalisation across grid sizes" paragraph was added immediately after the 5x5 results paragraph in Results. The paragraph states explicitly that: (a) demand N^* is the absolute number of origin-destination pairs sampled across the network, not normalised per intersection or per edge; (b) the 5x5 has 25 intersections vs 9 for the 3x3 (a factor of 2.78), so the same N^* corresponds to substantially lower per-intersection load on the 5x5; (c) trip lengths are longer on average on the 5x5 (mean trip ~ 3.3 blocks at 200 m each, vs ~ 2.0 blocks on the 3x3), so the per-intersection vehicle-time exposure is only partially compensated; (d) the traffic-generation process is otherwise identical (same SUMO seeds for each cell, same demand generation, same controller hyperparameters); (e) the 5x5 sweep used three seeds per cell. The paragraph cautions that the headline comparison in the present paper is PGQL versus max-pressure *on the same grid at the same N^** , and that cross-grid claims such as "PGQL gains more on the 5x5" should be read as the gain at matched absolute N^* , not at matched per-intersection load.

Comment 6. *The inclusion of a 5x5 network substantially strengthens the manuscript by demonstrating that the proposed controller is not limited to the original 3x3 benchmark. However, the revised experiments should not be interpreted as establishing scalability. Evaluating only two network sizes demonstrates feasibility on a larger network but does not characterize how controller performance, computational complexity, or learning behavior evolve with increasing network size. Consequently, any conclusions regarding scalability should be moderated to state that the proposed method was successfully validated on a second, larger benchmark network. A true scalability analysis would require evaluation across multiple progressively larger networks to establish performance trends as a function of network size.*

Response. All remaining "scalability" wording in the manuscript body and table captions was moderated to "feasibility" or "second-network validation." Specifically: the Results subsection title "5x5 grid scalability" is now "5x5 grid feasibility on a second network," the Network paragraph sentence "A larger 5x5 grid was also generated to test scalability" is now "A larger 5x5 grid was also generated for a second-network feasibility test," the Experimental Design phrase "The scalability sweep evaluated" is now "The second-network feasibility sweep evaluated," and the Table 3 caption "Scalability validation on a 5x5 grid" is now "Feasibility validation on a 5x5 grid." A grep of the resubmission DOCX returns zero hits for "scalability" or "Scalability" in manuscript body text or table captions. The Limitations section already lists networks of 49 or more intersections as untested, and the Future Work paragraph does not claim scalability and explicitly leaves a true scalability characterisation across multiple progressively larger networks as an open question.

Comment 7. *The manuscript compares PGQL with independently operating max-pressure controllers. An interesting question that remains unexplored is whether similar improvements could be achieved through hierarchical or regionally coordinated max-pressure control without reinforcement learning. Consequently, the observed performance gains should be interpreted as improvements over independent max-pressure rather than over all possible max-pressure-based architectures.*

Response. A new sixth bullet was added to the Limitations section in the Conclusion: "all comparisons in this paper are against independently-operating max-pressure controllers. Whether comparable performance gains could be obtained from a hierarchical or regionally-coordinated max-pressure architecture (without any reinforcement learning) was not tested. The observed gains should therefore be read as improvements over independent max-pressure, not as improvements over all max-pressure-based architectures." A corresponding bullet was added to the Future Work paragraph: "a comparison with hierarchical and regionally-coordinated max-pressure architectures, to establish whether the observed gains require reinforcement learning or could be achieved with classical multi-agent coordination alone."

Comment 8. *The manuscript demonstrates that a max-pressure warm start outperforms an actuated-controller warm start, thereby showing that the choice of expert influences the final learned policy. However, the manuscript does not include a cold-start (randomly initialized) PGQL baseline. Consequently, the experiments establish that one warm start is superior to another, but do not quantify the empirical benefit of warm-starting itself. A direct comparison between random initialization and max-pressure initialization would more convincingly demonstrate the practical value of the proposed warm-start framework in terms of convergence speed, training stability, and final performance.*

Response. This experiment was performed. A cold-start PGQL variant was built by setting `WARMSTART_BIAS = 0` (zeroing the initial bias toward the follow-max-pressure action). Both variants were trained on identical seeds for 10 episodes across the same four training loads used in the main paper, then evaluated at three demand levels with three seeds each. The results are reported as Table 7 in Results ("Cold-start ablation" subsection). The table is now preceded by a formal italic caption reading "Table 7. Cold-start ablation. Mean waiting time (mean $\pm$ sample SD across three seeds) for warm-started PGQL versus cold-started PGQL at three demand levels on the 3x3 grid. Cold-start beats warm-start at low and moderate load; warm-start retains a small edge at post-saturation." The finding is more nuanced than the paper's original design intuition suggested. After 10 training episodes, the cold-started variant matched or slightly outperformed the warm-started variant at low and moderate load (cold beat warm by 13.4 percent at $N = 200$ and by 5.9 percent at $N = 600$), while the warm-started variant retained a small 3.2 percent edge only at post-saturation $N = 1300$. The empirical value of the warm-start therefore is not a large permanent gap in final MWT. What the warm-start does contribute is (i) a lower starting-point MWT at training-episode zero (9.26 vs 11.07 s/car on $N = 200$, a gap of +1.81 s/car), and (ii) it avoids catastrophic early-training episodes: the cold-started agent produced MWT = 111.6 s/car at $N = 800$ in episode 3, versus 73.7 s/car for the warm-started agent, a 51 percent excess. The Discussion now describes the warm-start as providing safer early training rather than a permanent asymptotic advantage. The Limitations section notes that a longer-horizon comparison is needed to test whether the small post-saturation warm-start advantage grows or vanishes with more training.

Comment 9. *The algorithm neither prevents congestion, nor extends the capacity region until congestion occurs. It only improves performance during congestion. The Discussion and Conclusion would benefit from explicitly reconciling these observations so that readers understand that the demonstrated advantage is an incremental improvement in congestion management rather than a fundamental increase in network capacity or resilience.*

Response. Addressed by the same new Discussion paragraph cited under comment 4 above. The paragraph states directly that PGQL "does not raise N_{crit} appreciably above max-pressure in Table 1 and is not a fundamentally different congestion-management capability. The contribution is an incremental reduction in mean waiting time within the post-saturation regime, statistically significant under a paired Wilcoxon test, but bounded in magnitude (about 5 to 12 percent on the configurations tested)." The framing is now consistent across the Abstract (which speaks of "small but consistent improvements"), the Discussion, and the Conclusion.

The author thanks the reviewer for the level of detail in the refining comments. The three items the reviewer explicitly offered to defer (state-by-state decision-boundary audit, temporal-coordination versus residual-error causal analysis, and cold-start versus warm-start ablation) were performed and are reported in the revised Results section. All other comments were addressed by direct edits to the manuscript.

Thank you for addressing my comments.

The manuscript uses numerical citations; however, the numbering does not appear to follow Vancouver style. References should be renumbered so that citation numbers correspond to the order in which sources first appear in the text, and the reference list updated accordingly.

Response to Reviewer Comments

Manuscript: "Pressure-Guided Q-Learning: A Warm-Started Hybrid Controller for Signalised Urban Grids"

Author: Advik Singh

The reviewer's comment is quoted verbatim below and followed by a Response describing exactly how and where it has been addressed in the revised manuscript.

Comment 1. *The manuscript uses numerical citations; however, the numbering does not appear to follow Vancouver style. References should be renumbered so that citation numbers correspond to the order in which sources first appear in the text, and the reference list updated accordingly.*

Response. All in-text citations and the reference list have been renumbered in Vancouver style. Each reference is now assigned a number matching the order in which it first appears in the body of the manuscript, and the reference list is reordered accordingly. A first-appearance sweep of the revised body text produces the strictly sequential order 1, 2, 3, ..., 11. The full mapping from the previous numbering to the Vancouver numbering is:

- [2] -> [1] Wei, H., Zheng, G., Yao, H., and Li, Z. (2018). IntelliLight (KDD '18).
- [4] -> [2] Mannion, P., Duggan, J., and Howley, E. (2016). Adaptive RL for traffic signals.
- [10] -> [3] Haydari, A., and Yilmaz, Y. (2022). Deep RL for ITS: a survey. IEEE TITS.
- [11] -> [4] Noaeen, M., et al. (2022). RL in urban network traffic signal control: a systematic review.
- [1] -> [5] Varaiya, P. (2013). The max-pressure controller.
- [8] -> [6] Chen, C., et al. (2020). MPLight: decentralized deep RL for large-scale traffic signal control (AAAI).
- [3] -> [7] Wei, H., et al. (2019). PressLight (KDD '19).
- [9] -> [8] Zheng, G., et al. (2019). CoLight (CIKM '19).
- [5] -> [9] Watkins, C. J. C. H., and Dayan, P. (1992). Q-learning.
- [6] -> [10] Behrisch, M., et al. (2011). SUMO.
- [7] -> [11] Webster, F. V. (1958). Traffic signal settings.

The new numbering opens with the reinforcement-learning literature that motivates PGQL (IntelliLight [1], the Mannion adaptive-RL review [2], the Haydari IEEE TITS survey [3], and the Noaeen systematic review [4]), followed by Varaiya's max-pressure controller [5] as the classical baseline, then the larger-scale RL work (MPLight [6], PressLight [7], CoLight [8]), then the Watkins-Dayana Q-learning result [9] as the theoretical foundation, and finally SUMO [10] and Webster [11] in Methods. The math range [200, 1400] s in the Metrics subsection is unchanged because it is a mathematical interval rather than a citation.